# Information Systems as Infrastructure for University Research Now and in the Future

*Merrill Series on*
*The Research Mission of Public Universities*

A compilation of papers originally presented at a retreat
sponsored by The Merrill Advanced Studies Center
July 2012

Mabel L. Rice, Editor
Technical editor: Evelyn Haaheim

MASC Report No. 116
The University of Kansas

# TABLE OF CONTENTS

MASC Report No. 116

# Introduction

Mabel Rice

The Fred and Virginia Merrill Distinguished Professor of Advanced Studies and Director, Merrill Advanced Studies Center, The University of Kansas

The following papers each address an aspect of the subject of the sixteenth annual research policy retreat hosted by the Merrill Center: *Information Systems as Infrastructure for University Research Now and in the Future.*

We are pleased to continue this program that brings together University administrators and researcher-scientists for informal discussions that lead to the identification of pressing issues, understanding of different perspectives, and the creation of plans of action to enhance research productivity within our institutions. This year, the focus was on information systems and how they function as infrastructure in research universities currently, and what future trends in this infrastructure might be.

Our keynote speaker for the event, Dr David Shulenburger, discussed the present and future information infrastructure required for research success in universities, and the economic implications of that infrastructure.

Benefactors Virginia and Fred Merrill make possible this series of retreats: The Research Mission of Public Universities. On behalf of the many participants over more than a decade, I express deep gratitude to the Merrills for their enlightened support. On behalf of the Merrill Advanced Studies Center, I extend my appreciation for the contribution of effort and time of the participants and in particular to the authors of this collection of papers who found time in their busy schedules for the preparation of the materials that follow.

Twenty senior administrators and faculty from five institutions in Kansas, Missouri, Iowa and Nebraska attended the 2012 retreat. Though not all discussants' remarks are individually documented, their participation was an essential ingredient in the general discussions that ensued and the preparation of the final papers. The list of all conference attendees is at the end of the publication.

The inaugural event in this series of conferences, in 1997, focused on pressures that hinder the research mission of higher education. In 1998, we turned our attention to competing for new resources and to ways to enhance individual and collective productivity. In 1999, we examined in more depth cross-university alliances. The focus of the 2000 retreat was on making research a part of the public agenda and championing the cause of research as a valuable state resource. In 2001, the topic was evaluating research productivity, with a focus on the very important National Research Council (NRC) study from 1995. In the wake of 9/11, the topic for 2002 was "Science at a Time of National Emergency"; participants discussed scientists coming to the aid of the country, such as in joint research on preventing and mitigating bioterrorism, while also recogniz-

ing the difficulties our universities face because of increased security measures. In 2003 we focused on graduate education and two keynote speakers addressed key issues about retention of students in the doctoral track, efficiency in time to degree, and making the rules of the game transparent. In 2004 we looked at the leadership challenge of a comprehensive public university to accommodate the fluid nature of scientific initiatives to the world of long-term planning for the teaching and service missions of the universities. In 2005 we discussed the interface of science and public policy with an eye toward how to move forward in a way that honors both public trust and scientific integrity. Our retreat in 2006 considered the privatization of public universities and the corresponding shift in research funding and infrastructure. The 2007 retreat focused on the changing climate of research funding, the development of University research resources, and how to calibrate those resources with likely sources of funding, while the 2008 retreat dealt with the many benefits and specific issues of international research collaboration. The 2009 retreat highlighted regional research collaborations, with discussion of the many advantages and concerns associated with regional alliances. The 2010 retreat focused on the challenges regional Universities face in the effort to sustain and enhance their research missions, while the 2011 retreat outlined the role of Behavioral and Social sciences in national research initiatives.

Once again, the texts of this year's Merrill white paper reveal various perspectives on only one of the many complex issues faced by research administrators and scientists every day. It is with pleasure that I encourage you to read the papers from the 2012 Merrill policy retreat on *Information Systems as Infrastructure for University Research Now and in the Future.*

# Executive summary

## Information Systems as an Infrastructure for University Research Now and in the Future: Their Role in Helping Public Universities Cope with Financial and Political Stress

David Shulenburger, Senior Fellow, Association of Public and Land-grant Universities and Professor Emeritus, University of Kansas

- Data come at us rapidly from nearly every quarter and demand to be analyzed. High speed, high capacity computers, blindingly fast networks and massive storage capacity, huge quantities of digitally searchable text and considerable expertise will be required to deal with it all. And finally, it must be affordable. Can public universities afford to compete in this future?
- Public universities operate at a distinct disadvantage relative to our competitors in both private universities and in the corporate world. Public universities' major patrons, the 50 states, have been defunding them on a per student basis for last three decades.
- How do we as public universities acquire (or acquire access to) the information infrastructure that will enable us to maintain or improve our position in research, especially funded research, despite our financial weakness?
- What I recommend is that public universities promote the development of mechanisms and patterns of thought and behavior that enable sharing among all actors of the information infrastructure vital to the research enterprise.
- Elias Zerhouni's vision of the ideal future of medical research involved building the systems in which all research studies were placed freely available on-line. The scientific literature, genome, tissue and whole organism data sets and repositories and the research data bases that grew from his vision at NIH are precisely the kind of publicly available resources that enable public universities to compete and also permit all researchers, wherever they are housed, to be more productive.
- The information infrastructure required to compete in the research environment of the next decades will be less affordable to public research universities than to others because of their financial disadvantages. They would therefore differentially benefit if information infrastructure were made publically available to all without regard to financial factors. Making such resources available in this manner is good for science and for society.

## Trends, Disruption, and our Knowledge-Based Economy

Rob Duncan, Vice Chancellor of Research, University of Missouri

- The rate of advancement in our markets is accelerating today. Over the last twenty-five years, the primary seat of innovation and discovery have shifted from industrial laboratories to major, research-intensive universities, and hence social expectations are shifting to universities to lead future advances in technology commercialization that will preserve and extend the United States' international competitiveness. All advances in technology trigger creative disruption of pre-existing market structures, and universities are not historically good at managing such disruption.

- Twenty-five years ago, 70% of all R&D-100 awards were won by industrial laboratories, while today over 70% of these awards go to universities and public research foundations. Still, very little direct commercialization is conducted by universities, in favor of technology licensing of university technology to industry. Our future industrial competitiveness will continue to depend more on innovative methods of cooperation between universities and industry.
- Business today understands that the best ideas and technology rapidly displace earlier innovations in the markets, and while years ago many companies fought to extend their product life cycles by opposing technological advancement, today there is much more of a philosophy of embracing new ideas, and striving to be on the beneficial side of these inevitable disruptions. Similarly, universities that boldly innovate today, and which are open to new and much more aggressive strategies of technology management and industrial relations, will in general win big as well. Today, much more so than ever before, the risk of not taking the risk is actually much greater than the risk itself.
- History shows time and again that exceptional innovation moves disruptively against well-established trends, without attempting to be disruptive. Though it is easy for the common wisdom in Universities to think that small visionary efforts cannot create revolutionary advancement, in fact it is the only thing in retrospect that has. It remains critically important for university leaders to permit innovative faculty members to continue to do research that their peers often consider a clear waste of time.
- In many cases, innovations come in waves, with the much larger market penetration come only much later, when the fundamental organizational principles are discovered. Today universities can take advantage of these natural transitions through interdisciplinary innovation teams that critically evaluate and improve basic discoveries as they emerge, with a focus on how to scale upon the product demand to huge levels.

**Developing Infrastructure for Informatics Research: Experiences and Challenges**

Prem Paul, Vice Chancellor for Research and Economic Development, University of Nebraska-Lincoln

- Scientific advances are generating massive amounts of data, fostering the need for new strategies to access, store, analyze, and mine data. The need to manage "big data" is especially acute in the life sciences. t UNL, our faculty are working on plant and animal genetics and utilizing genomics in their research to improve productivity, particularly regarding traits for disease resistance and/or drought tolerance. These studies require major investments in computing and bioinformatics infrastructure and personnel trained in bioinformatics.
- Scientific advances are generating massive amounts of data, fostering the need for new strategies to access, store, analyze, and mine data. The need to manage "big data" is especially acute in the life sciences. UNL invested in 10Gb Dark Fiber connecting Lincoln and Kansas City at a cost of over $1 million to connect with Internet2. The cyberinfrastructure we developed to manage big data has been very helpful in supporting our faculty who require supercomputers.
- We have a bioinformatics service in our Center for Biotechnology core facility; however, the facility's staffing is not sufficient to meet the needs of all faculty because there are

more users than talented experts. We are exploring ways to add bioinformatics staff in the core facility and hire additional bioinformatics faculty, including a leader who can coordinate bioinformatics resources and services across campus.

- UNL has significant strengths in social and behavioral sciences, including the Gallup Research Center, the Bureau of Business Research, the Bureau of Sociological Research and the University of Nebraska Public Policy Center. Though each program is highly successful, there is an opportunity for strengthening these programs through collaborations. This is critical, especially considering the importance of social and behavioral sciences in major societal challenges. Therefore, we have launched a taskforce to better understand our institutional strengths and needs for infrastructure.
- UNL faculty members have recognized the need for a Research Data Center (RDC) to access economic, demographic, health statistics, and census data. RDCs provide secure access to restricted use of microdata for statistical purposes. Qualified researchers prepare proposals for approvals by the Census Bureau. Following approval, work is done in secure facilities where scientists may access centrally held data.
- Since RDCs are expensive to maintain and require hiring a director that is Census Bureau employee, it might be more appropriate to pursue a regional RDC that could serve universities in Nebraska, Iowa, Kansas, and Missouri. We propose to build such a center at UNL that would be available to our regional partners.
- Several years ago at the Merrill Conference, discussion took place regarding what can we do together that we cannot do alone – especially with regard to creating shared research infrastructure to support large-scale research projects and programs. The RDC concept represents such an idea for regional collaboration to access big data in social and behavioral science research.

**Skating to Where the Puck is <u>Going</u> to Be**

Steven Warren, Vice Chancellor for Research and Graduate Studies, University of Kansas

- What is it that we really want from information technology? Scientists want the speed and power to communicate, teach, and learn from anywhere in the world at any time, with ease. We want to have the power and speed to analyze remarkably complex problems.
- One of our ongoing goals is to transform the research administration experience for scholars at the University of Kansas by creating a fully integrated electronic research administration system - where researchers can check the financial balance of their grants, look at projections given their present rate of spending, update and submit a request to the Institutional Review Board, work on a new proposal, submit and monitor a travel request, and on and on.
- The most exciting place to be is on the front end of innovation. It can also easily be the most expensive, complicated, and disappointing place to be. Bottom line, letting others serve as the early adopters may mean that you get a better, more reliable and cheaper product in the end.
- A fundamental outcome of the IT revolution has been the change it has made in terms of how easy it is to collaborate with anyone almost anywhere. My experience has been that

scientists will seek out whoever they need to solve the challenges they face, often without regard to location. This has contributed to the explosion of scientific knowledge over the past couple of decades.

- It is getting much easier to detect certain types of scientific misconduct due to breakthrough technologies. The biggest change is in our ability to detect plagiarism. Scientific journals can now subscribe to services that will scan each submission they receive and compare it to countless related papers that have been published all over the world, in a search to detect instances of plagiarism.
- We live in truly remarkable times. The pace of technological and scientific innovation is staggering. Universities in fact are the origin of much of this disruptive, creative destruction that is rolling across virtually every corner of the world - but being part of the source of this revolution does not in any way inoculate us from its transformative effects. Will the basic model of research universities survive the exponential changes in information technologies?

**Information as a Paradigm**

Jeffrey Scott Vitter, Provost and Executive Vice Chancellor, University of Kansas

- We are in an information age. Computer science, information, and IT have made huge advances in the last few decades. Advances in computer technology have fundamentally changed the way we live. In fact, computer technology has become the infrastructure that drives commerce, entertainment, healthcare, national security, transportation and innovation.
- The takeaways that emerge about the growth of billion-dollar IT sectors: Each one of these sectors can trace its formation to university research and, in almost all cases, to Federally-funded research. It takes a long time for the research to pay off, in most cases one or two decades. Finally, the research ecosystem is fueled by the flow of people and ideas back and forth between university and industry. This system has made the United States the world leader in information technology.
- IT is not only a key driver for research at KU. We are also using IT in a broad sense to build an infrastructure for innovation – for instance, in our new Center for Online and Distance Learning (CODL), which helps faculty build online or hybrid courses and also serves as a central point for students to access online learning.
- KU was the first public university to adopt an open access policy, which makes faculty members' scholarly journal articles available online for free. KU Libraries hosts a public portal called KU ScholarWorks that provides free access to all faculty publications under the policy. KU is also using technology to maintain Faculty Professional Records Online (PRO), and to build capabilities for sophisticated analytics that allow us to examine our effectiveness and productivity as a university.
- The state of Kansas demands a flagship university in the top tier of public international research universities. KU continues to actively engage with communities throughout Kansas and the world, with a focus upon entrepreneurship, commercialization of technology, and vibrant business partnerships. All of these depend upon IT.

- Though it is no longer so relevant that Kansas is at the geographic and geodesic center of the continental United States, it is significant that through IT, we can truly immerse ourselves anywhere in the world, link together key partners, and form vibrant collaborations. IT drives society, and it drives KU. Through our research at KU and as a core part of how we operate, we use information in fundamental ways to improve our understanding of the world and to make it a better place.

**Scholarly Communication in the Age of New Media**

Brian Foster, Provost, University of Missouri

- **S**cholarly communication is critical for both the research and educational missions of universities. Given the centrality of scholarly communication to the mission of higher education, it is unsettling that all we know about the current model is that it will not work in the future. Most of the focus of this paper is on publication. The main point is that we are really looking beyond books and journals as we know them, to media of the future that are not known.
- Faculty members expect open communication with regard to content in scholarly publication. However, there are many limitations on open communication that arise from security and commercialization interests such as protecting intellectual property (IP), national security issues, and classified research.
- The main quality assurance mechanism for scholarly publishing is peer review. We expect that work archived in prominent publications has been vetted and can be considered reliable. However, since the most respected publishing venues tend to be conservative, it raises the question of whether really innovative, high-impact or interdisciplinary research will be "accepted" for publication.
- In many ways, the most important function of scholarly publication is the archival function. Yet there are some significant questions about the long-term viability of the digital publications as an archival function. For example, there does not seem to be a coherent plan to migrate the staggering amounts of "archived" digital data if significant new technologies emerge.
- Finally, articles and monographs are critical resources for advanced education. But there are many practical and legal challenges to the use of such materials in digital format. It is a certainty, at least in my view, that we will see dramatic changes in the formats of educational materials. These issues will create daunting questions with respect to IP rights of faculty creating the materials.
- Whatever the future path, there will be unexpected consequences. For instance, there is a perspective by which digital subscriptions in libraries limit access to scholarly results. One doesn't need to "sign in" to take a paper volume off the shelf in the Library; but if one wants access to the digital journals, one has to be a "member" of the "organization" (e.g., university) to be able to "sign in" to get access to the on-line material—a condition of the licensing.
- The issues involving scholarly communication are very complex. A key issue is that we must identify the unintended consequences of change—and the only certainty is that dramatic change will occur. A sustainable revenue stream must be found. In any case, we

must mitigate the unintended consequences such as limiting access as a condition of library subscriptions. We need new models for scholarly communication, and we need to understand that the only thing we really know is that the current system is not sustainable.

**Smart Infoware: Providing University Research Stakeholders Soft Power to Connect the Dots in Information Haystacks**

Chi-Ren Shyu, Director, Informatics Institute, University of Missouri

- While resources have been allocated to build computing infrastructure everywhere in the nation, the value of infoware to assist the university scientific communities has been underestimated, or even ignored. The organization and coordination of available infoware is needed to leverage regional talents to equip researchers with soft power as opposed to the hardware-based computing muscles.
- The informatics community at MU continuously develops infoware to serve the worldwide research community in three major areas – bioinformatics, health informatics, and geoinformatics. This infoware has played a significant role in handling the ever- growing size of data in genomics, proteomics, biometrics, and imaging technologies.
- Without smart infoware to analyze the data, the financial burden to purchase larger computer clusters becomes bigger and unmatchable to the growth of information. All the infoware shares the same goal: to provide open-source tool access to the entire scientific community with speedy search from large-scale and complex data sets that normally cannot be organized and processed without high performance computing.
- Infoware has been developed independently by colleagues in Iowa, Kansas, Missouri, and Nebraska. However, most informaticians are unaware of these developments in their surrounding institutions. Thus it is unlikely to expect researchers in other fields to understand the regional talents that can greatly enhance their research using the existing infoware. Therefore, it is necessary for infoware developers from the region to meet and put together an info-warehouse for tool sharing and education.
- Moreover, a research social network which is searchable by university researchers and industry partners is also needed for the region. This linkage of researchers may consist of co-authored publications, collaborative proposals for extramural grants, student committee memberships, national/international committee services, etc.

**Finding Subatomic Particles and Nanoscopic Gold in Big Data**

David Swanson, Director, Holland Computing Center, University of Nebraska

- Modern data intensive research is advancing knowledge in fields ranging from particle physics to the humanities. The data sets and computational demands of these pursuits put ever-increasing strain on University resources. The need to constantly evolve and grow resources economically requires careful strategy and provides incentive to combine and share resources wherever possible.
- HCC maintains a Campus Grid that is able to transparently submit large batches of jobs first to the campus, then to Big 10 peers Purdue and Wisconsin, and finally to the OSG. This method of computing is also able to access resources on the Amazon EC2 Cloud. Recent studies by the DOE concerning storage have continued to conclude that Cloud re-

sources, even if capable of the scales demanded by data centric science -- not yet confirmed -- are currently far too expensive compared to locating resources at Universities and Labs.

- HCC's relationship with the Open Science Grid (OSG) has grown steadily over the last several years, and HCC is committed to a vision that includes grid computing and shared national cyberinfrastructure. The experience with OSG to date has proven to be a great benefit to HCC and NU researchers, as evidenced by the HCC's usage of over 11 million cpu hours in the last calendar year.
- HCC is a substantial contributor to the LHC (Large Hadron Collider) experiment located at CERN, arguably the largest cyberinfrastructure effort in the world. HCC operates one of the seven CMS "Tier-2" computing centers in the United States. Tier-2 centers are the primary sites for user analysis of CMS data, and they are also the primary sites for generating collision simulations that are used by physicists throughout the experiment; these jobs are submitted over the grid by a central team of operators. The Nebraska Tier-2 currently has 3000 processing cores and hosts 1200 TB of CMS data with redundant replication.
- Movement of data sets of these sizes requires the use of high performance networks. UNL has partnered with other regional Universities in the Great Plains Network (GPN) to share the cost of obtaining and maintaining connectivity to Internet2. The Tier2 site at HCC routinely moves data at rates approaching 10 GigaBits per second to and from other LHC collaborators, often via this shared Internet2 link.
- While shared high throughput computing (HTC) is a major thrust at HCC, it should be emphasized the majority of computing done at HCC is still traditional high performance computing (HPC), since this is what most HCC researchers currently need to advance their research. The fact that HPC coexists so well with HTC at HCC is strong evidence this model of shared computing can be extended to other locations.

**On the End of Paper Based Communication**

Perry Alexander, Director, Information and Telecommunication Technology Center (ITTC), University of Kansas

- The post-PC world is upon us. We are consuming more information in more ways than ever in our history. The industrial revolution snatched up our information infrastructure and made its products commodities. Yet, Accounting still needs physical receipts for my trips. Whatever happened to the paperless office promised two decades ago? The technology is most clearly here and available. What's missing? What does paperless mean?
- Paperless means literally less paper, lowering costs, and producing less waste. Instead of reams of paper and stamps, we now consume terabytes of bandwidth and storage. We need greater bandwidth and storage, new kinds of data archives, more software with far higher complexity. Still, paper establishes trust. We need *new ways of establishing trust* that reflect our new ways of storing and transmitting information.
- The tools of virtual trust are cryptographic functions for encrypting and signing messages and cryptographic protocols for exchanging information. Encrypting information with a key provides the same trust as a sealed. Signing information with a key provides the same

trust as a physical signature. Protocols provide us methodologies for using encryption and signing to exchange information.

- Asymmetric key cryptography gives us the tools to electronically replace envelopes that guarantee confidentiality and physical signatures that guarantee integrity. *Protocols* then specify how those tools are used in practice. We are seeing these protocols implemented in everything from software distribution systems to lab information maintenance.
- Establishing trust electronically is problematic. Key management – particularly revocation of compromised keys – is an ongoing area of research and development. But the tools are there for us to use. The time is now for us to move forward and begin to put trust on equal footing with information in the electronic world.

**The Role of Information Systems in Clinical and Translational Research (Frontiers: The NIH Clinical and Translational Science Award Driving Information Systems)**

Paul F. Terranova, Richard J. Barohn, Lauren S. Aaronson, Andrew K. Godwin, Peter Smith, and L. Russ Waitman, University of Kansas Medical Center

- Headquartered at the University of Kansas Medical Center, the NIH-CTSA-supported Frontiers program, more formally called The Heartland Institute for Clinical and Translational Research, is a network of scientists from institutions in Kansas and the Kansas City region. The Frontiers program is developing more efficient use and integration of biorepositories, genomic information and biomedical informatics which are important components for attaining the CTSA goals.
- Many partners are involved with accomplishing the goals of our Frontiers program. These partners include academic institutions and health care institutions and centers in the region. The basic infrastructure of Frontiers includes the following components: Clinical Research Development Office, Clinical and Translational Research Education Center, Biomedical Informatics, Biostatistics, Clinical and Translational Science Unit (CTSU), The Institute for Advancing Medical Innovations, Translational Technologies Resource Center, Pharmacokinetics/Pharmacodynamics (PK/PD), Personalized Medicine and Outcomes Center, Pilot and Collaborative Studies Funding, Community Partnership for Health, Regulatory Knowledge and Support, and Ethics.
- The bio-repository includes numerous components requiring integration of multiple sources of information flow with a goal of enhancing discovery to improve health. Our goal is to enter genome and other types of 'omic' data such as metabolome, transcriptome and proteome into the system.
- Bioinformatics services at KUMC are provided largely by the Smith Intellectual and Developmental Disabilities Research Center which is a collaborative effort with KU at Lawrence Bioinformatics studies and utilizes methods for storing, retrieving and analyzing biological data, such as DNA, RNA, proteins and their sequence as well as their structure, function, pathways and interactions. The overall mission of the core facility is to advance the understanding of integrative functions in biological systems, including human, through the application of computational models and data analysis with focus on Next Generation Sequencing Data Analysis and Microarray Data Analysis. The core identifies opportunities and implements solutions for managing, visualizing, analyzing, and interpreting genomic data, including studies of gene expression, pathway analysis, protein-

DNA binding, DNA methylation, and DNA variation, using high-throughput platforms in both human and model organisms.

**Research, Data, and Administration Management in the Animal Health/One Health Plan**

James Guikema, Associate Vice President for Research, Associate Dean of the Graduate School, Kansas State University

- The theme of the current Merrill Conference – information systems as infrastructural priorities for university research both now and in the future – unites the research office and the information technology (IT) office at each university institution within our four-state region.
- There is an unprecedented demand on our IT infrastructure from nearly all sectors of our collective clientele. Students are sophisticated users of IT networks, and push the limits of what we can provide in communication, academic course content, and social networking. This is amplified when considering the needs of the distance-education arena. Our research-faculty investigators are developing larger databases that challenge our ability to archive, manage, manipulate, mine, and share essential information.
- High on the list of research priorities are the 'omics' – genomics, proteomics, metabolomics, lipidomics, etc. – with the i5K genome project [sequencing and annotating 5,000 insect genomes], described elsewhere in this volume, as an excellent example. Policy makers, and the managers of funding programs at both the state and federal levels, are sending our universities mixed messages regarding what data may (or must) be shared and what data must be secured.
- Relative to IT demands, our universities are looking into the tunnel at the light – not knowing if that light is in fact the end of the tunnel, or if it is the headlight of the oncoming train. One thing is certain: a sense of a deadline approaching. This was perhaps best described by Dr. Samuel Johnson: "Nothing so concentrates the mind as the sight of the gallows," and the IT challenges are relevant across all disciplines on our campuses.
- Each of our institutions has risen to the IT challenge in various areas and disciplines. This article seeks to place the IT / research infrastructure challenges within a 'comparative medicine' context, since these challenges touch upon research strengths of the region, of strengths within each institution, and of the growing regional opportunity represented by the relocation of a major federal comparative medicine laboratory [NBAF] to Manhattan, KS.

**Trends in Technology-enabled Research and Discovery**

Gary K. Allen, CIO, University of Missouri-Columbia; Vice President for IT, University of Missouri System

James Davis, Vice Provost for IT & CIO, Iowa State University

- Discovery, innovation, learning, and engagement are core to the mission of the university, and support for those objectives pervades all processes. An effective information technology infrastructure (or *cyberinfrastructure*) is central to the success of a robust research program. The infrastructure must provide anywhere, anytime access to information, peer collaborators, systems, and services needed to advance the program.

- Even with the rapidly increasing capacity of contemporary storage, the management of large collections of data is demanding. The complexity of maintaining large data stores coupled with curation requirements and rapidly expanding security requirements makes a compelling case for developing an institutional approach to data management. Another challenging component of processing large research data sets is simply moving them quickly and reliably.
- IT is in the largest outsourcing trend in history, and public cloud Infrastructure as a Service (IaaS) will be a key component of outsourcing decisions because it offers commoditized infrastructure and increased agility. A different rate of evolution is evident in IT services supporting research activities. New models are rapidly developing among communities of use whose members have these requirements in common. Project teams expect that technology will mitigate the impact of distance and create a community where participants can interact as if they were located in the same space.
- Joint efforts, many of which are regionally focused, are also growing in number as institutions work together to develop new approaches to satisfy their research cyberinfrastructure needs. In many cases, these efforts include both research universities and corporate partners.
- We believe that there is an important role for a regional approach to developing strategies to bridge between the campus and national research cyberinfrastructure initiatives. All of the represented states in the Merrill retreats are EPSCoR-eligible; this could represent a significant opportunity to obtain federal funding support to begin creating a regional support model for cyberinfrastructure.

**Communicating Science in an International Arena: the i5K Initiative**

Susan Brown, University Distinguished Professor of Biology, Kansas State University

- As technical advances lower the costs of high through-put sequencing, genome sequencing is no longer limited to large sequencing centers and external sources of funding. This creates special challenges in how to store, share and analyze huge datasets with an international cohort of collaborators.
- As sequencing costs plummet, sequencing projects are directed toward more ambitious goals. At K-State, we established a center for genomic studies of arthropods affecting plant, animal and human health. The next generation sequencing technologies that have yielded the most dramatic cost reductions produce exceedingly large datasets. Many algorithms have been developed to assemble a genome from these whole genome shotgun reads, and all are computationally expensive.
- In March 2011, the i5k initiative was announced. The goal of the i5k is to sequence the genomes of 5000 insect and related arthropod species. The i5k consortium is interested in all insects of importance to agriculture, medicine, energy production, and those that serve as models, as well as those of importance to constructing the Arthropod Tree of Life.
- Sequencing 5000 insect genomes will provide a wealth of information about the largest group of species on earth. The genome sequences will serve as a resource for gene discovery. Genome sequences will also serve as substrates for analysis, to detect signatures of selection and structural rearrangements associated with their evolutionary history. Genome

sequences will also serve as gateways for biological inquiry, providing a "parts list" of genes for investigation.

- At K-State, we are using NGS sequence data to improve the original *Tribolium castaneum* genome assembly. In addition, we have sequenced the genomes of three related species. Each project generates terabytes of data to be archived, analyzed, and shared with an international group of collaborators.
- We currently provide community access to three genome projects at K-State through agripestbase (agripestbase.org). To solve problems in distributed resource sharing we are working with our high performance computing (HPC) group to explore solutions offered by Globus. Sequencing the genomes of the interacting members of these communities will provide the foundation for arthropod community genomics and even larger datasets to consider.

**Advancing Clinical and Transformational Research with Informatics at the University of Kansas Medical Center**

Lemuel Russell Waitman, Gerald Lushington, Judith J. Warren, University of Kansas Medical Center

- Biomedical Informatics accelerates scientific discovery and improves patient care by converting data into actionable information. Pharmacologists and biologists receive improved molecular signatures; translational scientists use tools to determine potential study cohorts; providers view therapeutic risk models individualized to their patient; and policy makers can understand the populations they serve. Informatics methods also lower communication barriers by standardizing terminology describing observations, integrating decision support into clinical systems, and connecting patients with providers through telemedicine.
- The University of Kansas Medical Center's (KUMC) pursuit of a National Institutes of Health (NIH) Clinical and Translational Science Award (CTSA) catalyzed the development and integration of informatics capabilities to specifically support translational research in our region. The NIH-CTSA-supported Frontiers program, also called The Heartland Institute for Clinical and Translational Research (HICTR), is a network of scientists from institutions in Kansas and the Kansas City region.
- The vision for the informatics section is to provide rich information resources, services, and communication technologies across the spectrum of translational research. Broadly the initiative would adopt methods which facilitate collaboration and communication both locally and nationally, convert clinical systems into information collection systems for translational research, and provide innovative and robust informatics drug and biomarker discovery techniques.
- In addition, the informatics section will work with state and regional agencies to provide infrastructure and data management for translational outcomes research in underserved populations, and measure clinical information systems' ability to incorporate translational research findings.
- In the latter years of the grant we plan to leverage regional informatics capabilities to complement the long history of engaging the community through outreach via telemedi-

cine, continuing medical education, and our extensive community research projects in urban rural and frontier settings. In the coming year, Frontiers plans to optimize the use of clinical systems for disseminating translational evidence and recording measures to verify adoption.

**Creating Information Infrastructure for Research Collaboration**

Arun K. Somani, Anson Marston Distinguished Professor, Electrical and Computer Engineering, Iowa State University

- High performance computing (HPC) is essential for advanced research in virtually all disciplines of science and engineering, and in particular in the fields of bio-, materials-, and information-technologies, all identified as strategic priorities of Iowa State University.
- The merit of the new HPC platform includes the far-reaching and impactful research growth that it enables, by accelerating knowledge and discovery, spanning areas as diverse as animal sciences, plant genomics, climate modeling, and wind power generation.
- ISU has large, well established interdisciplinary research and education programs at the forefront of biological sciences. These research efforts span microbial, plant and animal species, and are fully integrated with teams of computational scientists due to the transformation of biology as a data-driven science.
- The computational and bioinformatics tools needed to drive such research share common methodologies and computing needs. The emergence of high-throughput DNA sequencing technologies and their rapid proliferation and throughput gains during the past five years has created a compelling need for the HPC equipment.
- A group of faculty members was identified to develop external funding for this proposal, specifically targeting the NSF MRI (major research instrumentation program. A successful $2.6M proposal for a large heterogeneous machine was developed that met the needs of a plurality of researchers.

# Information Systems as an Infrastructure for University Research Now and in the Future: Their Role in Helping Public Universities Cope with Financial and Political Stress

**David Shulenburger,** Senior Fellow, Association of Public and Land Grant Universities and Professor Emeritus, University of Kansas

Data come at us rapidly from nearly every quarter and demand to be analyzed. High speed, high capacity computers, blindingly fast networks and massive storage capability, huge quantities of digitally searchable text and considerable expertise will be required to deal with it all.

According to George Dyson, the data universe grows at 5 trillion bytes a second. Just as a seemingly inert lump of matter disguises the intricate physics and geometry within it, the flood of data contains patterns that we will never know of unless we have the capacity to fully understand them. Where did the stock market's "flash crash" of May 2010 come from? Does "junk" DNA actually play a key role in biological functioning? Do patterns of word usage and grammar from the 1700s suggest deeper understanding of the physical world than we heretofore imagined?

Ubiquitous connectivity, inexpensive sensors, falling storage costs have made the analysis of "big data" mandatory. Examples of big data sets that may reveal both surface and hidden meanings to us are:

- Routinely captured and shared medical records data bases
- Genome: for millions of people at an increasingly modest cost
- Human Biome: 10,000+ organisms inhabit the human body
- Culturomics: Over 15 million books have been digitized. Word frequency, birth and propagation of ideas, celebrity, demonization, taboos, commerce, business cycles, etc., etc., etc.
- Google with 10 billion web pages, e-mail, texts, blogs, etc.
- Facebook and other "social media"
- Securities transactions, on-line financial transactions

The future requirements for an information infrastructure to support research within a university are not hard to discern and are reasonably well understood. In the listing below, I add only one novel characteristic, affordability.

I come from a long history of public

**Characteristics of Information Infrastructure Required For Research Success in the Future**

- High Speed Transmission
- High Speed/Capacity Computing
- Massive Storage Capacity
- Fully Digitally Searchable Text and Data
- Content! Content! Content!
- Sophisticated expertise needed to take full advantage of all this.

- **Affordable**

university involvement, deep at one university, the University of Kansas, and broad, involving all the 217 public research member universities and systems that belong to the Association of Public and Land-grant Universities (APLU). The perspective brought by knowledge of that set of universities causes me to take this paper in a direction that recognizes their economics. And that leads me to ask "Can public universities afford to compete in this future?"

**The Declining Economic Fortunes of Public Universities**

Public universities operate at a distinct disadvantage relative to our competitors in both private universities and in the corporate world. Mark Emmert, then President of the University of Washington, and a keen observer of the economic damage being done to the "system" of public universities independently funded by the fifty states observed: *"[public research universities have] evolved into a critical national asset…we could see 50 different independent decisions made in 50 different states and the result would be a national tragedy!"* Former President of the University of Vermont, Mark Fogel, on examining the same terrain, chillingly summed up the feeling of major public university presidents, "We are haunted by the specter that our enterprise has seen its best days . . ."

Briefly, the primary financial problem of public universities is that their major patrons, the 50 states, have been defunding them on a per student basis for last three decades as Figure I aptly illustrates. While public universities have been able to replace lost educational appropriations with added tuition dollars, the source from which funds are derived has subtle effects on how they are spent by university officials. I shall have more on the effects of source on distribution later. Figure I includes funding data for all public higher education from community colleges to the most sophisticated graduate universities. Segregating the data for the public universities in the Carnegie classifications "very high research universities" and "high research universities" in Figure II demonstrates that both categories of research universities have suffered from the cuts. Since 2007 real appropriations have fallen at least 10% for both categories.

I suspect that it will be quite some time, if ever, before state per student funding of public universities returns to the levels of two decades ago. There are many reasons for this judgment but a sufficient reason is that the competition for scarce state resources has grown very intense. A major claimant for such resources in every state has been the federally-mandated state share of Medicaid funding. Figure III shows that state funding for Medicaid and higher education were close to equal in 1987 but Medicaid expenditure has grown by a factor of 5.3 in the intervening years while higher education grew by only 1.3.

Higher education has a very big stake in the 2012 presidential election, as the debate concerning public funding of health care, illustrated in Figure III, makes clear. Other competitors for state funding include prisons, pension funding and infrastructure. Many believe that the growing private goods nature of higher education has led state legislators to conclude that it should be funded by students rather than taxpayers. Whatev-

er the argument, few informed observers believe that state subsidy of higher education will grow quickly in the near future.

**The Growing Funding Advantage of Private Universities**

While state funding has declined, tuition at public research universities has increased but only by enough to slightly more than offset the loss of appropriated funds. Private research universities charge significantly larger tuitions than public universities and have far larger endowments per student. The result is that their available funds per student are far greater than those of public universities. As a result private research universities spend more than 2.3 times as much per FTE student on educational and general expenditures as do public universities. That multiple has been maintained over time (Figure IV).
This higher level of expenditure permits private universities to acquire the quality and volume of resources they desire. Such resources include faculty and staff. As a result faculty salaries at public research universities have fallen from near parity with those at private research universities in the 1970s to around 80% of their level today (Figure V).

Similarly, private research universities have increased their instructional staffing per student more than public universities. In 2007 the level of staffing per students at private research universities reached 120% of the level of public research universities (Figure VI).

The relatively smaller funding of public research universities in and of itself reduces their ability to compete with the privates. But the increasing proportion of funding that comes from tuition in the publics and the decreasing proportion coming from tuition in private institutions exert subtle effects on the choices administrators make about where to spend institutional resources (Figure VII).

As tuition increases as a proportion of E&G expenditure, relatively more resources are spent on activities that are seen as directly benefiting undergraduate students and relatively less on activities that benefit graduate students or research. A prime example of this can be observed in expenditures on research libraries. The library budget as a percent of total university E&G falls more in public than in private universities as the proportion of the education and general budget derived from tuition increases. This effect probably generalizes to other areas of the budget, including research.

Another explanatory factor for the difference in level and trend of E&G expenditures on libraries is the differential in academic program makeup at public and private research universities. Public universities have approximately 21% of their students in graduate programs while private universities have 49%. Given the greater need for access to the research portion of the collection by graduate students, it is to be expected that private universities would spend the larger proportion of their funds on libraries. Similarly, the pressure from undergraduates to spend monies on programs directly benefiting undergraduate students would be greater at public research universities. Thus, the "tune" is called not only by the increasing proportion of the budget derived from tuition, but from the relative pro-

portions of graduate and undergraduate students (Figure VIII).

Another source of relative disadvantage comes from the larger subsidy that public universities provide to federal research than do private universities. This probably arises because public universities understand that they are at a disadvantage relative to private universities in the competition for federal grants and subsidize their research activities more in order to make their proposals relatively more attractive to funding agencies (Figure IX).

The strategy apparently works, for public universities have gained an increasing share of federal research grants over time as the subsidy for research they conduct was increased. But the burning question is whether subsidy of any size can compensate in the long-term for the public university's rapidly falling real state subsidy and increasing funding from a budgetary source that is less compatible with funding of research (Figure X).

**Growing Research Funding Success of Industry and Government Labs**

But both private and public universities have competition for funding from other sources. Industry, government labs and non-profits each have increased their share of the federal research dollar since 2002. Thus, our thinking about public university competitiveness for research has to be broadened beyond the private university competitors (Figure XI).

**A Plausible Empirical Explanation for the Growing Disadvantage of Public Research Universities**

Is there a coherent argument that explains why public university subsidy from the states has fallen and why more of the research dollar is flowing away from universities? I think there is and it has recently been articulated in an article by Gordon Gauchata of the University of North Carolina.[1] Gauchata examines polling data of the American citizenry and finds:

- Public trust in science has not declined since the 1970s except among conservatives and those who frequently attend church,
- That there is negligible evidence for the cultural ascendency thesis, which suggests that trust in science will increase over time. Nor do poll results support the alienation thesis that predicts a uniform decline in public trust in science,
- Polling results are consistent with claims of the politicization thesis and show that conservatives experienced long-term group-specific declines rather than an abrupt cultural break,
- Educated conservatives uniquely experienced the decline in trust in science. (Figure XII)

Gauchata found that the public defines "what science is" in three distinct ways:

(1) As an abstract method (e.g., replication, empirical, or unbiased);

(2) As a cultural location (e.g., takes place in a university or is practiced by highly credentialed individuals); and

(3) As one form of knowledge among other types such as commonsense and religious tradition.

Conservatives were far more likely to define science as knowledge that should conform to common sense and religious tradition. When examining a

series of public attitudes toward science, conservatives' unfavorable attitudes are most acute in relation to government funding of science and the use of scientific knowledge to influence social policy. Conservatives thus appear especially averse to regulatory science, defined here as the mutual dependence of organized science and government policy

Gauchata draws on the analysis of Lave, Mirowski and Randals[2] to tie this research to the rise of neoliberal science management regimes since 1980, particularly the insistence on the commercialization and privatization of knowledge that has created substantive shifts in the organization and practice of science. Perhaps the most obvious shift is the rollback of government funding for, and organization of, public research universities. (Underlining is mine) He continues that the changes in the organization of science include:

(1) Increased government outlays to private corporations rather than universities;

(2) Intellectual property rights restricting public access to scientific knowledge; and

(3) Reversal of the postwar trend of viewing teaching and research as mutually reinforcing activities.

Gauchata's work neatly fits with the facts I observe. Federal research money is flowing differentially toward private corporations. The State subsidy to public universities has been reduced over time. The amendment of the copyright law overtime clearly has had the effect of restricting public access to scientific knowledge. Finally, the conservative resistance toward openly teaching some scientific findings are particularly relevant when we consider global climate change—and growing public skepticism toward that problem-- or when we consider the development of genomics and its implications for private interests.

I do not exclude explanations other than Gauchata's for the state of public universities and the distribution of research funding. I cite his explanation because it is consistent with the facts I observe. It seems to me that universities are relatively powerless to change the balance of power in governmental bodies between conservative and less conservative members; we must work with what we have. The relevant question is, "How do we as public universities acquire or acquire access to the information infrastructure that will enable us to maintain or improve our position in research, especially funded research, despite our financial weakness?"

**What Can be Done to Improve the Research Competitiveness of Public Research Universities?**

Many of the techniques used to enhance university research competitiveness are beyond the scope of this inquiry, which is limited to questions of information infrastructure. Clearly, public universities can and should continue proven techniques such as focusing their research investment in promising areas of research so that their grant funding possibilities will be enhanced and hiring accordingly.

Similarly, implementing each of the ten recommendations of the National Research Council's June 2012 report, Research Universities and the Future of America, would clearly improve the prospects of public universities. Implementation of recommendations 1, 2, 6, 7

and 10 would reduce or help public universities better cope with the financial disparities between public and private universities but implementation of all 10 recommendations would be of much benefit to all research universities. None of the recommendations directly address the information infrastructure topic of this inquiry but several of them would provide additional funding or administrative flexibility that might address problems in this area (Figure XIII).

**Public Access to Information Infrastructure as a Means of Preserving and Strengthening Public Research University Competitiveness**

What I recommend to address the information infrastructure needs is for public universities to promote the development of mechanisms and patterns of thought and behavior that enable sharing among all actors of the information infrastructure vital to the research enterprise. I am not advocating that public research universities form a collective and share among themselves. If they were to form a collective that excludes others, they would ensure only a shared poverty, not access to sufficient information infrastructure resources to make them competitive. Many of the information infrastructure resources public universities need are available to those who have infrastructure wealth, the private universities, private enterprise and the federal government.

Interestingly, what I spell out here as good for public research universities is also good for society. No less a visionary than Isaac Newton saw this: *"If I have seen further it is by standing on the shoulders of giants."*[3] Newton's giants reside in public universities and private universities throughout the world as well as in government laboratories, the private labs of individual tinkerers and in industry. Leading the effort to keep the "shoulders of giants" available as a platform for all despite the resources that are available is essentially the role I foresee for public universities. They will be a major beneficiary.

But directly to the relevant point, the research of Furman and Stern into factors that led to increased citation, patents and diffusion of research reached this conclusion, "Overall, the ability of a society to stand on the shoulders of giants depends not only on generating knowledge, but also on the quality of mechanisms for storing, certifying and accessing that knowledge."[4] The mechanisms for storing and accessing knowledge are the core of information infrastructure.

Elias Zerhouni's (former NIH Director) vision of the ideal future of medical research involved creating the conditions/building the systems in which all research studies, all genome structures, all chemical information, all data sets, etc., were placed freely available on-line so that connections that an individual scientist may never encounter by reading the literature become discoverable.[5] He fervently believes that such a system would dramatically enhance the productivity of science and set about creating a set of systems at NIH that would produce that end. The scientific literature; genome, tissue and whole organism data sets and repositories; and the research data bases that grew from his vision at NIH are precisely the kind of publicly available information infrastructure resources that enable public universities to

compete with private universities, government labs and private industry and that also permit all researchers, wherever they are housed, to be more productive.

Finally we must recognize that the massive amount of knowledge that we have amassed and are adding to daily requires special storage and access conditions if they are to be truly available to researchers. Robert Merton foresaw this nearly 50 years ago when he said, *"Perhaps no problem facing the individual scientist today is more defeating than the effort to cope with the flood of published scientific research, even within one's own narrow specialty.*[6]" The thousands of articles that might have been available on a given subject then are an exponential multiple of that number now. Unless scientific literature, patents, data sets, etc., are stored in a digital form, accessible and readable by computers, they are not truly accessible.

The major forms of public access discussed below are:

I. Public Access To Scholarly Research
   A. Public Access to Federal and Foundation Funded Research
   B. Public Access Deposit Requirements by Universities
   C. Open Access Journals
   D. Digital Repositories by Universities, Disciplinary Societies
II. Open Research Data
III. Open Research Organisms and Materials
IV. Open Access to High Speed/ Capacity Computing
V. Open Access to High Speed Networks

## I. Public Access to Scholarly Research

### IA. Public Access to Federal and Foundation Funded Research

Public access to the scientific literature has grown significantly in recent years. Figure XIV lists the major initiatives in the world by funders and national governments to provide access. Some provide access immediately upon publication in a journal and others provide access after a delay of up to one year. The largest depository is NIH's PubMed Central. Should the Federal Research Public Access Act (FRPPA) be enacted by Congress, publications arising from research funded by all federal agencies that fund more than $100 million per year in research would ultimately become available to the public for free (Figure XIV).

The volume of material accessed through PubMed is quite significant. The users come from the academy, industry and the general public (Figure XV).

The call to permit public access to articles that otherwise would be available only to those with subscriptions or in institutions with subscriptions has been intense and enduring. The greatest concern express by scholarly journals has been that making their articles public, even after a lag of 12 months, would damage their revenue streams. The NIH repository has been in operation over five years and during that time no major scientific journal has gone out of business or reduced the number of issues or articles published. Financial analysts that follow commercial publishers have concluded that NIH policy has not caused a substantial number of journal cancellations. The good health of the scientific scholarly journal market is unfor-

tunately evident as seen by the fact that STEM journal prices keep going up.

Physics is the field with longest OA archiving tradition dating back to 1991. Most physicists report that they go first to arXiv to find their scientific literature, not to the journals. Despite this access preference, the major physics journals say no journal cancelations have resulted from this archive that contains the most important contributions to physics; two physics journal publishers even have their own mirror versions of arXiv.

**IB. Public Access Deposit Requirements by Universities**

Fifty-two research funders worldwide mandate that researchers publicly archive publications arising from their funding. Figure XVI lists those funders.

Universities have begun to "mandate" that their faculty members publicly archive their journal research publications. While "mandate "is the term utilized, it is a misleading term because the universities participating have created mandates as a result of faculty action, not by administratively dictating faculty action. This is strong evidence that faculty are beginning to understand the need for such openness. In the US, the first university-wide mandate was by the faculty at MIT, followed closely by the faculty at the University of Kansas and at Duke. Harvard University is often thought of as being the first U.S. University to create a mandate but that mandate was not university-wide, instead it was by the faculty of Arts and Sciences. Since then faculty of other colleges at Harvard have issued similar mandates but at this writing the mandates do not cover all faculty at Harvard. Worldwide, 149 universities have mandates in place. Other universities in the U.S. and throughout the world have such mandates under consideration.

The first institutional mandate was not by universities or funding agencies but by the U.S. government. Its mandate is different from that of universities as it does not require that publications of federal employees be placed in a repository, but it forbids those employees from giving exclusive copyrights to publishers of their work (see Title 17 United States Code section105). The federal government is ultimately the copyright holder of their work and has the right to publish that work itself even if it has been published in a scholarly journal. Thus the federal government could "publish" all works of employees by placing them in a publicly accessible repository if it chose to do so.

**IC. Open Access Journals**

A powerful way of creating public access is for the journal of publication to make all of its articles available to the public for free. As of July 2012 there are 7,902 scholarly journals that follow this practice. The most prominent journals in this group are those published by the Public Library of Science where the editorial boards of the journals frequently have multiple Nobel Prize winners serving on them. Some of the journals are financed by requiring authors to pay publication fees, others are financed by scientific institutions and some are financed by donated funds. The number of OA journals is growing at a very rapid rate (Figure XVII).

**ID. Digital Repositories by Universities, Disciplinary Societies**

The most prominent way to make researchers' scholarly works available to

the public is for disciplinary groups or universities to create digital archives into which such works can be voluntarily deposited. Figure XVIII lists the numbers of such archives that have been created by country and by discipline. Such archives are generally available to search engines, making works placed in them easily available world-wide. In addition to materials voluntarily placed in such archives, some categories of materials such as master's and doctoral theses may be placed there as a result of a university mandate.

**The Advantage to the Author of Open or Public Access:** Regardless of the vehicle that makes an article publicly accessible, the robust finding is that articles that are publicly accessible are generally cited more frequently than those that can be accessed only through subscriptions. In a major review of 31 studies, Alma Swan writing in 2010 found that 27 of the 31 found a citation advantage for Open Access articles.[7] Interestingly, physics articles that are placed in arXiv are cited five times more frequently than those that are not and 20% of the citations actually occur before the article appears in a journal.[8]

There is also evidence that practitioners and even members of the general public wish to access scientific studies and do access them when they have the opportunity to do so. For example, mental health practitioners sent links to articles variously open access and gated. One week later the open access article was found to have been read twice as often.[9] Six out of ten physicians change their initial diagnosis based on information accessed on line.[10]

## II. Open Research Data

While there is not widespread agreement among individual researchers that research data should be shared, there is considerable U.S. and international official recognition that it should be. For example, the OECD in supporting open research data concludes that Open Research Data reinforces Open Scientific Inquiry by:

- Encouraging diversity of analysis and opinion
- Promoting new research & making possible testing of new or alternative hypotheses and methods of analysis
- Supporting studies on data-collection methods and measurements
- Facilitating the education of new researchers
- Enabling the exploration of topics not envisioned by the original investigators
- Permitting the creation of data sets when data from multiple sources are combined.[11]

According to Britain's Royal Society, "the potential of the Internet to facilitate collaboration among both professional and amateur scientists may pave the way for a second open science revolution as great as that triggered by the creation of the first scientific journals." They argue that this highly desirable end is dependent on routine publication of datasets in intelligible, assessable and usable formats which it calls "intelligent openness." Such publication would:

- allow a new breed of data scientists to search for unsuspected relationships, such as between disease

mechanisms and the properties of drug-like compounds.
- improve the detection of scientific error and,
- help to build public trust in science in areas such as climate science and genetic modification.

Finally and controversially, they argue that this sharing of data sets would be facilitated if high-quality and publicly accessible datasets were to be given as much credit as standard publications in Britain's research excellence framework.[12]

Both the NSF and NIH require researchers they fund to release research data to other researchers (details in Figure XIX). Their requirements lack definition and enforcement at this point but since the requirements are backed by law, one can be sure enforcement eventually will follow.

The development of norms among individual researchers that they should share research data not developed with federal funding is spotty. IPSCR at the University of Michigan is a respected repository for social and political scientists to share data and other such vehicles exist. Most journals have a policy about data sharing to verify research results and some journals simply require that the data on which an article is based be deposited with them.

A growing body of research demonstrates that publicly accessible data bases produce the same sorts of results one gets from publicly accessible journal articles, i.e., they are used more. For example, Heidi Williams considered the follow-on results from the Human Genome Project's open human genome vs. the Celera Corporation's closed one. Celera sequences were available only to those who paid for them, but Human Genome Project's sequences were publicly available. The Celera-sequences genes led to about 30% fewer articles about genotype-phenotype links than did the Human Genome projects, demonstrating that public data is used more. She found similarly reduced numbers of diagnostic tests based on Celera articles and increased numbers based on the HGP.[13]

## III. Open Research Organisms and Materials

Use of genetically identical animals and materials with common properties is essential for generalization of research. Some such organisms/materials are available from open sources maintained by governments/ foundations/ universities and others are available from sources that see them as a revenue source. Just as with scholarly research and data, researchers find that organisms/materials from open sources produce greater benefits than those from closed sources. Two examples:

***Mice from Open source vs. Closed source mice***

- Research based on open mice generated substantially more follow-on research and greater "horizontal extension" of follow-on research.
- Research based on open mice is more likely to be found in applied research journals suggesting that the open mice research might lead to faster commercialization.[14]

***Biological Resource Center materials vs. Commercial Sourced materials***

- Articles based on openly accessible materials got 220% more citations

than those based on commercially sourced materials.

- When materials moved from private archives (closed source) to BRCs (open source), citation rates to articles increased 50 to 125%.[15]

**IV. High Speed/Capacity Computing**
**V. High Speed Networks**

I will say little about these two because little needs to be said. Clearly without the super-computer centers funded by NSF over several decades, all but the wealthiest universities would have been without access. While the cost of high speed computing has fallen massively, the need for greater speed and capacity continues to present itself. Problems like global modeling of climate demand more and more capacity and speed. The availability of publicly accessible cutting edge computing will permit low resourced universities to continue to be competitors in all areas of research.

Similarly, Federal funding agency support for high speed networks enabled university research. The development of various communal university institutions continues to permit both public and private universities to enjoy access to what is now a necessity. Such "public access" must continue in the future or the more poorly funded actors will be unable to compete.

**Concluding Comments**

My argument is simply that the information infrastructure required to compete in the research environment of the next decades will be less affordable to public research universities than to others because of their financial disadvantages. They would therefore differentially benefit if information infrastructure were made publically available to all without regard to financial factors. Making such resources available in this manner is good for science and for society.

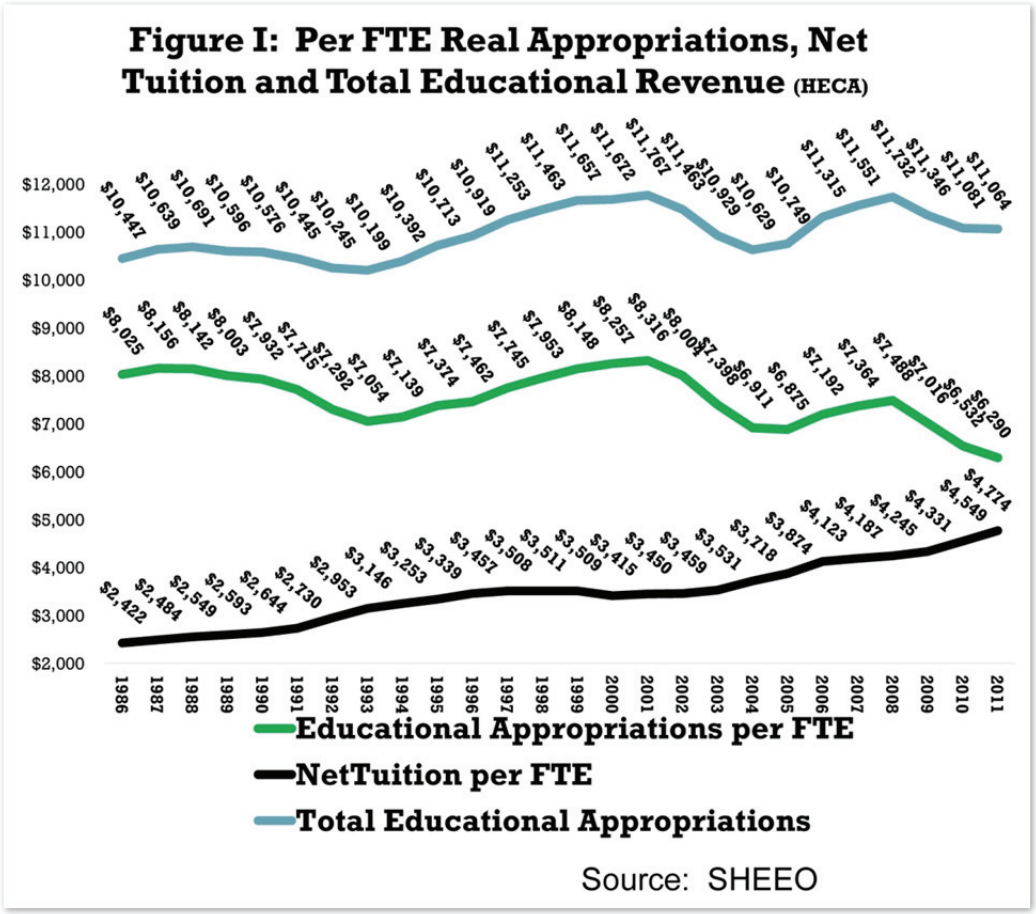
From this argument I proceeded to examine the state of availability of three elements of the information infrastructure: the scholarly literature, research data and research organisms and materials. Much progress has been made in making these information infrastructure elements publicly available. Scientific gain has resulted; much more progress is needed. It is clearly in the interest of public research universities to advocate for that progress and to act on their own campuses to make the elements under their control publicly available.

**References**

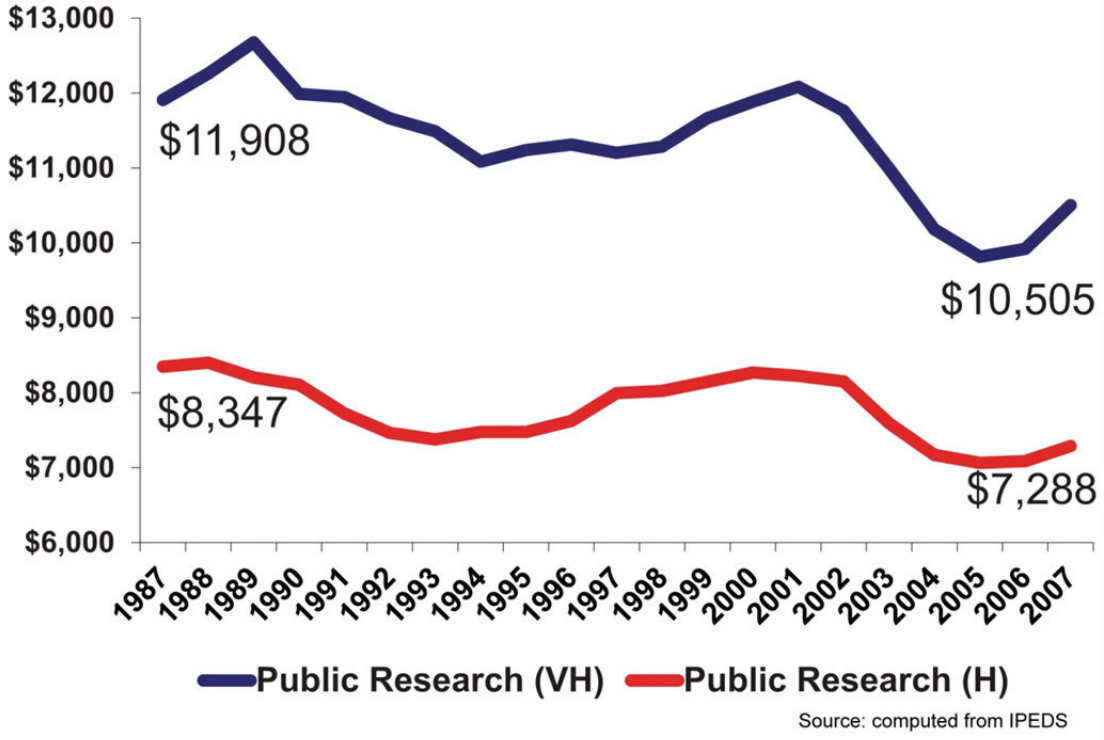

1. "Politicization of Science in the Public Sphere, A Study of Public Trust in the United States", 1974 to 2010 Gordon Gauchata, University of North Carolina, Chapel Hill, American Sociological Review April 2012 vol. 77 no. 2 167-187
2. Lave, Rebecca; Mirowski, Philip; Randalls, Samuel. 2010. "Introduction: STS and Neoliberal Science." Social Studies of Science 40:659–75.
3. Isaac Newton, Letter to Robert Hooke, February 5, 1675
4. Jeffery L. Furman and Scott Stern, "Climbing Atop the shoulders of Giants: The Impact of Institutions on Cumulative Research" The American Economic Review 101 (5), 2011
5. Barbara J. Culliton, Published online before print March 2006, doi: 10.1377/hlthaff.25.w94"Extracting Knowledge From Science: A Conversation With Elias Zerhouni" Health Affairs, May 2006 vol 25 no 3 w94-w103
6. Robert K. Merton, "The Matthew Effect in Science," 159 Science 56-63 (1968).
7. http://eprints.ecs.soton.ac.uk/18516/2/Citation_advantage_paper.pdf, 2010
8. Gentil-Beccot, Mele and Brooks, CERN, 2010
9. Hardesty Journal of Clinical Psychology (2008)
10. Letter of Francis Collins at http://publicaccess.nih.gov/Collins_reply_to-Pitts121611.pdf

11. 2007 OCED Recommendations Concerning Public Access to Research Data from Public Funding
12. "Science as an Open Enterprise: Open Data for Open Science," http://www.insidehighered.com/news/2012/06/22/royal-society-wants-britain-see-equal-value-datasets-and-journal-articles#ixzz1yWkXHwAz
13. Heidi Williams. NBER working paper #16123
14. Murray. et.al. in The Rate and Director of Inventive Activity, Univ. of Chicago Press (forthcoming 2012)
15. Furman and Sterns, American Economic Review 08/2011

# Figures:

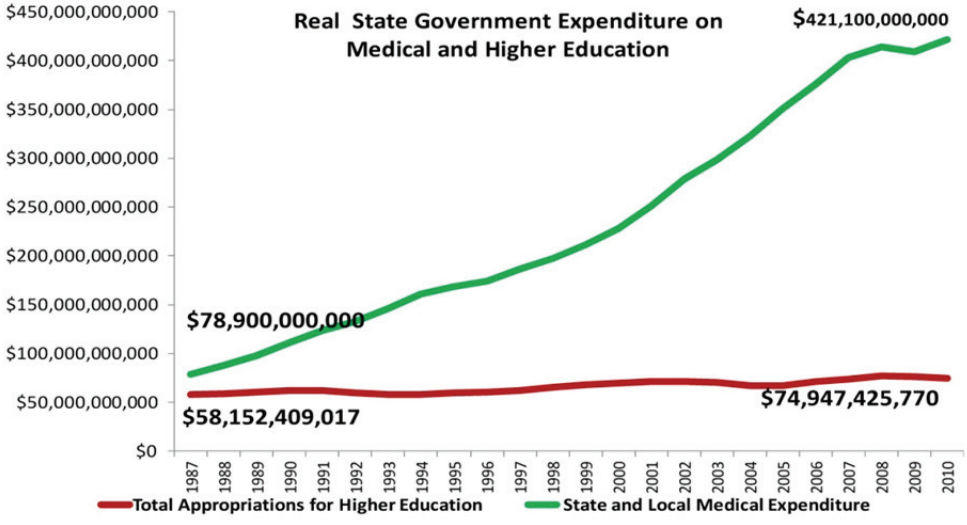


## Figure II: State and Local Real, Per Student Appropriations



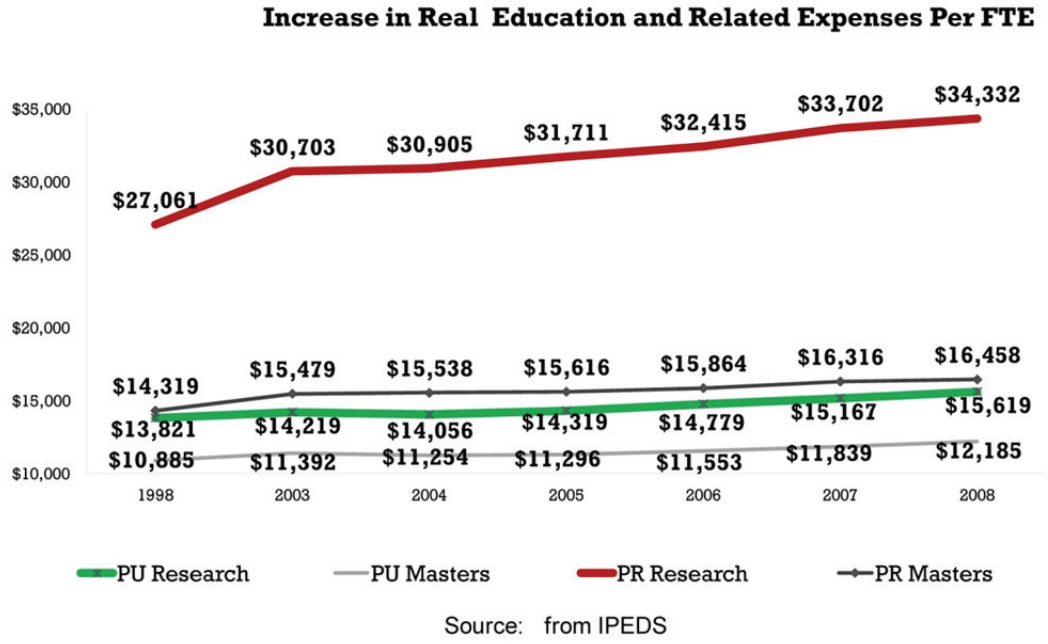

Source: computed from IPEDS

## Figure III: The Competition for State Funding is Fierce



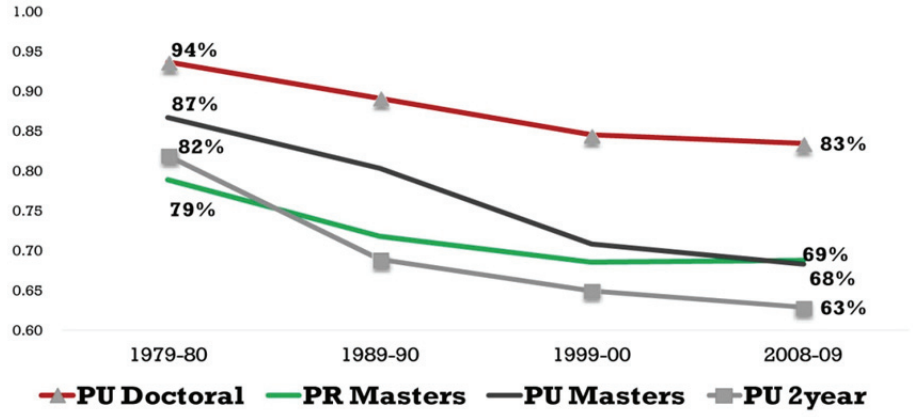

source: Government Spending.com

## Figure IV: Public Universities spend less than half as much per FTE student as private research universities.



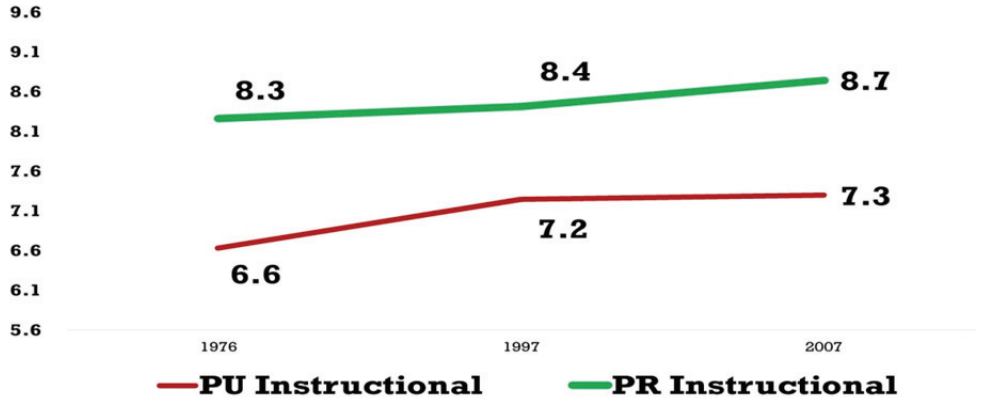

Source: from IPEDS

## Figure V: Average 9-month Salaries of full-time Faculty in Various Carnegie Categories as a Ratio of those in Private Doctoral Universities



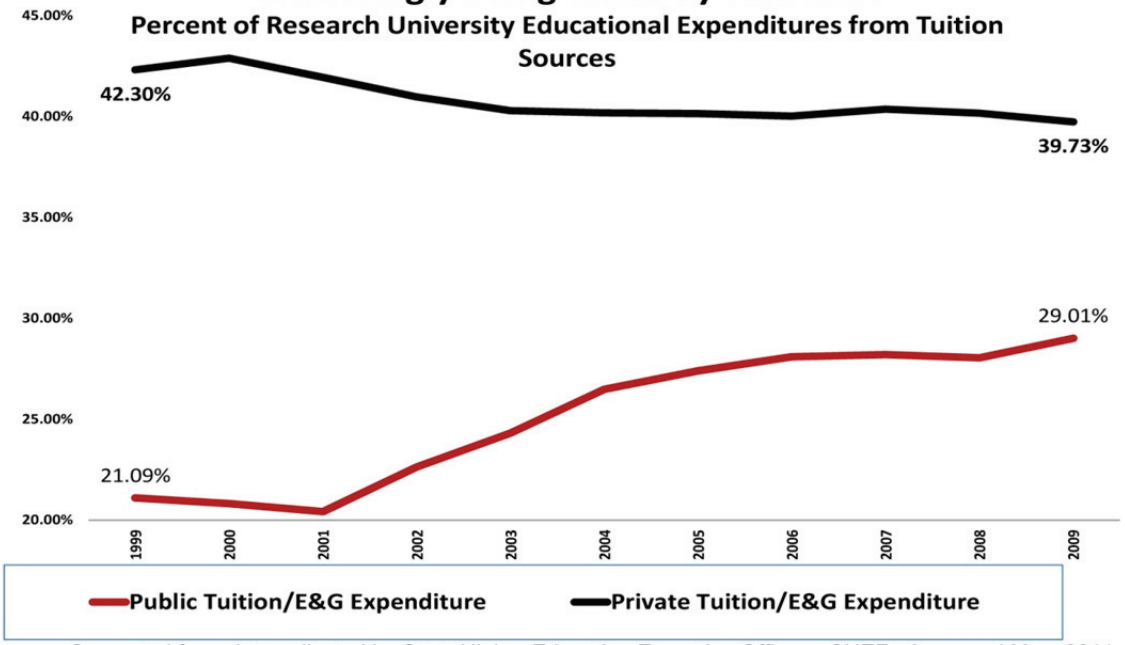

Source: NCES, *The Condition of Education 2010*, Indicator 44.

## Figure VI: Ratio of Instructional Staff per 100 Students at Public and Private Universities



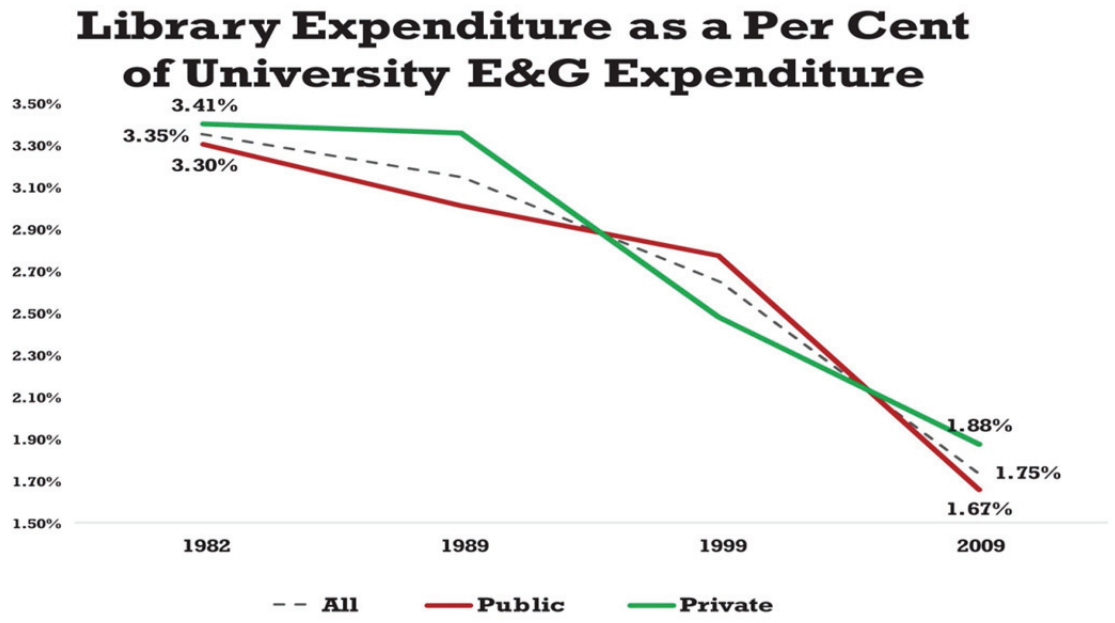

Source: NCES, *Digest of Education Statistics 2009*, Table 244.

## Figure VII: The Tune at Public Universities is increasingly being called by students.



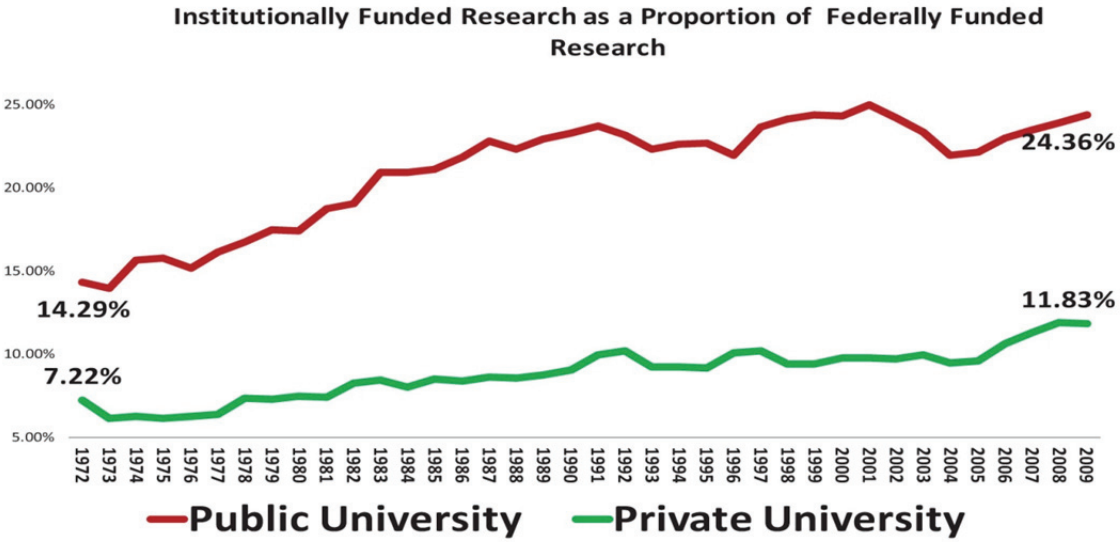

Source: Computed from data collected by State Higher Education Executive Officers, SHEF. Accessed May, 2011.

Figure VIII



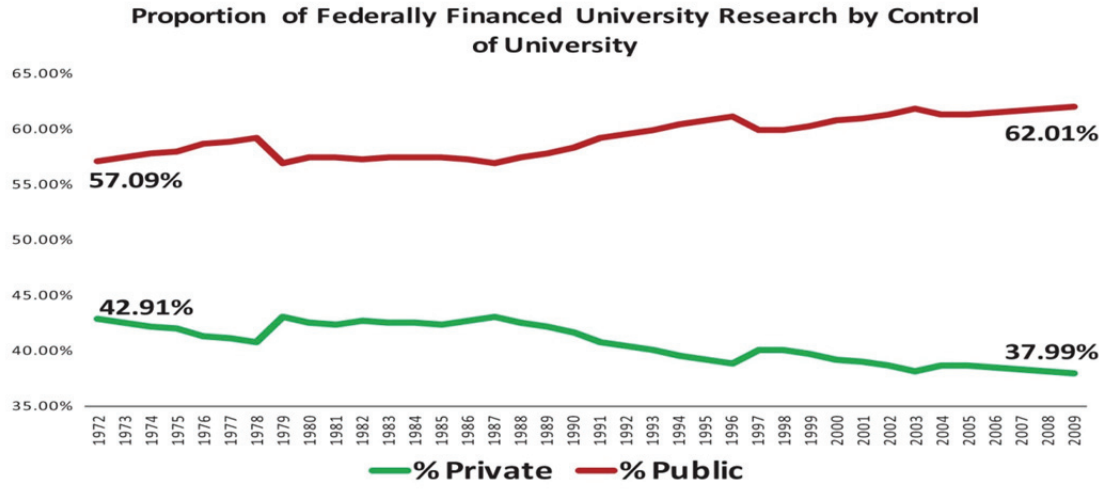

Source: Computed from ARL data base on Member Library Funding

**Figure IX: The academy has steadily increased the subsidy it provides for research with its own funds and public universities subsidize research more heavily than do private universities.**



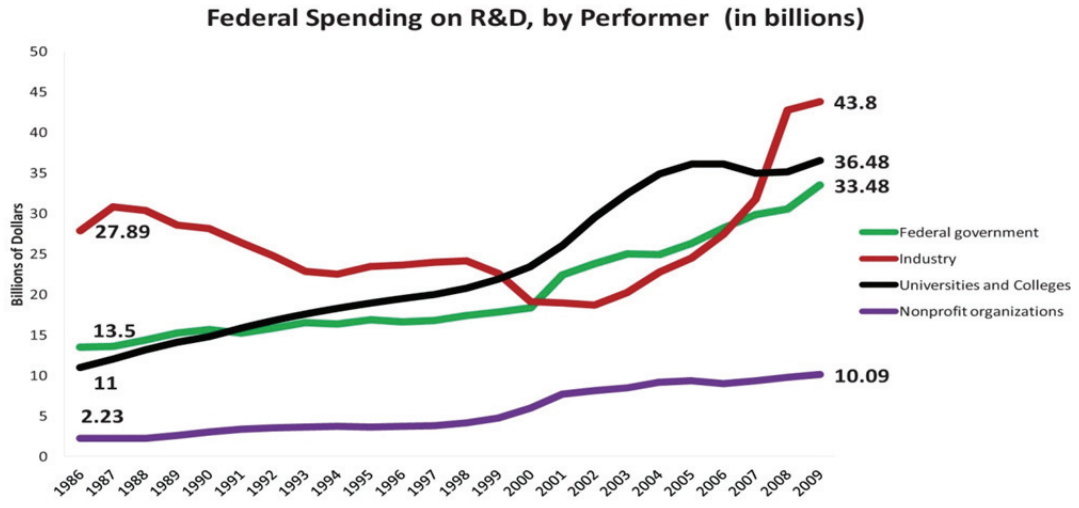

source: NSF Research Expenditures

**Figure X: Despite the adverse developments, public universities have maintained their competitiveness for federal grant awards but given the increasing discrepancy in faculty salaries, staffing levels and the ever larger subsidy of external research, can this success continue?**



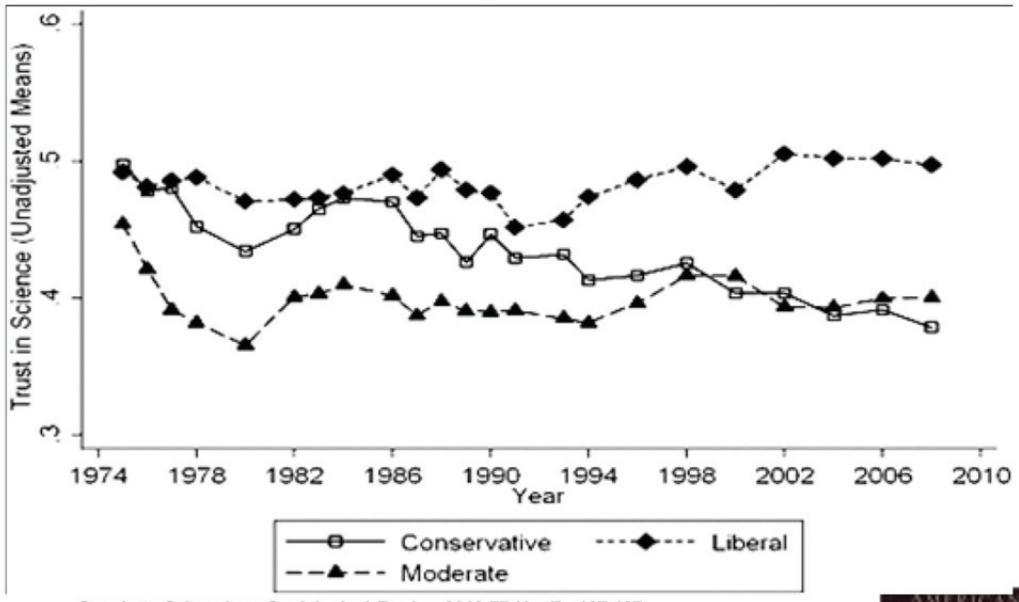

Source: NSF Research Expenditures

**Figure XI: Since 2000 Federal Expenditure on R&D Performed by Industry has far Outpaced that Performed by all Others**




Science and Engineering Indicators: 2012 http://www.nsf.gov/statistics/digest12/portfolio.cfm#4

**Figure XII: Unadjusted Means of Public Trust in Science for Each Survey Year by Political Ideology**




Gauchata G American Sociological Review 2012;77 (April) :167-187

## Figure XIII: NRC Research Universities and the Future of America Recommendations

1. The federal government should adopt stable and effective policies, practices, and funding for university-performed R&D and graduate education.
2. States should provide greater autonomy for public research universities so that these institutions may leverage local and regional strengths to compete strategically and respond with agility to new opportunities.
3. Increase university cost-effectiveness and productivity.
4. Strengthen the business role in the research partnership, facilitating the transfer of knowledge, ideas, and technology to society, and accelerate "time-to-innovation."
5. Create a Federal Strategic Investment Program that funds initiatives at research universities critical to advancing education and research in areas of key national priority.
6. The federal government and other research sponsors should strive to cover the full costs of research projects.
7. Federal and State governments should reduce or eliminate regulations that increase administrative costs, impede research productivity, and deflect creative energy without substantially improving the research environment.
8. Improve the capacity of graduate programs to attract talented students by addressing issues such as attrition rates, time-to-degree, funding, and alignment with both student career opportunities and national interests.
9. Secure for the United States the full benefits of education for all Americans, including women and underrepresented minorities, in science, mathematics, engineering, and technology.
10. Ensure that the United States will continue to benefit strongly from the participation of international students and scholars in our research enterprise.

## Figure XIV: Status of Public Access

- European Commission is funding Publishing and the Ecology of European Research (PEER) to investigate effects of depositing authors' peer-review manuscripts in a freely accessible location. Action in 2012
- British Government announced intention in Dec. 2011 to require all public funded scientific research be published in OA Journals
- Howard Hughes, Wellcome Trust require deposit of articles
- NIH has required deposit in PubMed Central since April, 2008, of articles emerging from research they support. Contains 21 million citations. 26,000 articles deposited each month.
  - 930 journals automatically provide full content
  - 300 additional journals provide final , published NIH articles
  - 1620 journals provide individual articles.
- FRPPA is now in Congress with 35 bi-partisan congressmen signed on. Principal Republican co-sponsor is Kevin Yoder of Kansas. FRPPA would expand an NIH-like requirement to all federal funding agencies with more than $100 million annually in research grants

## Figure XV: The Volume of Access due to PubMed is Huge

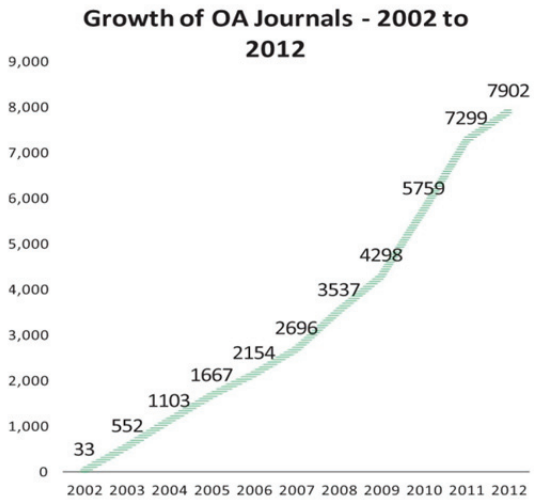
- Pub Med access is 500,000 on typical weekday.
- 1 million+ articles downloaded each day
  - 25% visitors from academic institutions
  - 40% visitors from general public
  - 17% visitors from companies

## Figure XVI: Funder Mandates – 52 World-Wide

Agency Nationale de la recherche (ANR) (Humanities and Social Sciences Branch) (29 Jul 2008)
Arthritis Research UK (04 Jan 2007)
Arts and Humanities Research Council (AHRC) (06 Sep 2007)
Australian Research Council (06 Dec 2006)
Autism Speaks (13 Nov 2008)
Biotechnology and Biological Sciences Research Council (BBSRC) (23 Jul 2006)
British Heart Foundation (BHF) (09 Jan 2007)
Canadian Breast Cancer Research Alliance (CBCRA) (02 Oct 2007)
Canadian Breast Cancer Research Alliance (CBCRA) (28 Jun 2009)
Canadian Cancer Society (CCS) (01 Oct 2008)
Canadian Health Services Research Foundation (CHSRF) (23 Apr 2009)
Canadian Institutes of Health Research (CIHR) (04 Jan 2007)
Cancer Research UK (09 Jan 2007)
Chief Scientist Office (Scottish Executive Health) (CSO) (09 Jan 2007)
Department of Health (UK) (DoH) (09 Jan 2007)
EUR-OCEANS Consortium on Ocean Ecosystem Analysis (21 Mar 2011)
Economic and Social Research Council (ESRC) (23 Jul 2006)
Engineering & Physical Sciences Research Council (EPSRC) (04 Oct 2011)
European Commission - 2 (EC) (20 Aug 2008)
European Research Council (ERC) (04 Jan 2007)
Fonds de la recherche en sante Quebec (FRSQ) (05 Feb 2009)
Fonds zur Foerderung der wissenschaftlichen Forschung (FWF) (09 Oct 2006)
Government of the Principality of Asturias (06 Feb 2009)
Health Research Board (Ireland) (HRB) (16 Feb 2008)
Heart and Stroke Foundation of Canada (05 Jul 2010)
Howard Hughes Medical Institute (HHMI) (02 Oct 2007)
Hungarian Scientific Research Fund (OTKA) (22 Jul 2009)
Institute of Education Sciences (25 Jun 2009)
International Development Reseach Centre (IDRC) (13 Jul 2011)
Irish Higher Education Authority (HEA) (22 Aug 2008)
Irish Research Council for Science, Engineering & Technology (IRCSET) (18 Sep 2007)
JISC (Joint Information Systems Committee) (02 Oct 2007)
Madrid Autonomous Community of Spain (CM) (28 Mar 2009)
Medical Research Council (MRC) (23 Jul 2006)
Michael Smith Foundation for Health Research (MSFHR) (23 Aug 2009)
National Health and Medical Research Council (NHMRC) (09 Dec 2006)
National Health and Medical Research Council (NHMRC) (22 Feb 2012)
National Institutes of Health (NIH) (25 Dec 2007)
National Research Council (NRC) Canada (16 Jul 2008)
Natural Environmental Research Council (NERC) (16 Aug 2006)
Norwegian Research Council (05 Feb 2009)
Ontario Institute for Cancer Research (OICR) (27 Jun 2008)
Research Foundation Flanders (FWO) (07 Apr 2007)
Science Foundation Ireland (SFI) (18 Feb 2009)
Science and Technology Facilities Council (STFC) (20 Oct 2006)
Spanish General State Administration (16 May 2011)
Swedish Research Council Formas (25 May 2010)
Swedish Research Council Vetenskapradet (15 Oct 2009)
Swiss National Science Foundation (SNF) (10 Aug 2007)
Telethon Italy (01 Mar 2010)
The Dunhill Medical Trust (26 Jul 2011)
Wellcome Trust (23 Jul 2006)

source: http://roarmap.eprints.org/

## Figure XVII: 7902 OA Journals in the World (6/25/2012), Wide-spread Geographically, In all Disciplines




Countries with more than 100 OA Journals

| | |
|---|---|
| United States | 1231 |
| Brazil | 743 |
| United Kingdom | 543 |
| India | 414 |
| Spain | 414 |
| Egypt | 315 |
| Germany | 248 |
| Canada | 238 |
| Romania | 228 |
| Italy | 210 |
| Turkey | 198 |
| Colombia | 183 |
| France | 154 |
| Iran | 141 |
| Chile | 139 |
| Poland | 133 |
| Argentina | 124 |
| Mexico | 120 |
| Australia | 119 |
| New Zealand | 109 |
| Switzerland | 106 |
| Japan | 106 |

Source: www.doaj.org/

## Figure XVIII: Repositories are Distributed World Wide and Across Many Disciplines

| | # of Institutional Repositories |
|---|---|
| US | 388 |
| UK | 206 |
| Japan | 137 |
| China | 33 |
| India | 54 |
| Australia | 48 |
| Germany | 149 |
| France | 66 |
| Brazil | 62 |

| | # of Disciplinary Repositories |
|---|---|
| Biology | 98 |
| Chemistry | 62 |
| Math | 67 |
| Physics | 62 |
| Medicine | 195 |
| CS/IT | 113 |
| Geology | 114 |
| History | 170 |
| Business/Econ | 130 |
| Education | 106 |
| Multidisciplinary | 1,339 |

http://www.opendoar.org/

**Figure XIX: NSF and NIH "Require" Data Sharing**

- As required by 45 CFR part 74.36, grantees that are institutions of higher education, hospitals, or non-profit organizations must release research data first produced in a project supported in whole or in part with Federal funds that are cited publicly and officially by a Federal agency in support of an action that has the force and effect of law
- NSF Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing. **NSF Data Management Plan Requirements -** Proposals submitted or due on or after January 18, 2011, must include a supplementary document of no more than two pages labeled "Data Management Plan". This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results.
- NIH believes that data sharing is essential for expedited translation of research results into knowledge, products, and procedures to improve human health. NIH endorses the sharing of final research data to serve these and other important scientific goals and expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers. "Timely release and sharing" is defined as no later than the acceptance for publication of the main findings from the final data set. All investigator-initiated applications with direct costs of $500,000 or more (excluding consortium F&A costs) in any single year are expected to address data-sharing in their application.

# Trends, Disruption, and our Knowledge-Based Economy

**Robert Duncan,** Vice Chancellor for Research,
University of Missouri

The rate of advancement in our markets is accelerating today. Over the last twenty-five years, the primary seat of innovation and discovery has shifted from industrial laboratories to major, research-intensive universities, and hence social expectations are shifting to universities to lead future advances in technology commercialization that will preserve and extend the United States' international competitiveness. All advances in technology trigger creative disruption of pre-existing market structures, and universities are not historically good at managing such disruption. This paper addresses and quantifies these trends through case studies, and then discusses the strategies and structures that we have put in place within the University of Missouri for consideration as a method for management of technical innovation, entrepreneurialism, and the associated creative disruption.

**Introduction**

Most scientific and technical innovation today occurs in universities. Twenty-five years ago, 70% of all R&D-100 awards were won by industrial laboratories, while today over 70% of these awards go to universities and public research foundations. Nonetheless, very little direct commercialization is conducted by universities, in favor of technology licensing of university technology to industry. Our future industrial competitiveness will continue to depend more on creating innovative methods of cooperation between universities and industry, and there is opportunity now in innovating new management procedures that permit closer cooperation that bridge across the public and private sectors.

While these changes are themselves very disruptive at many levels, our markets today are becoming more and more accustomed to coping in a continuous state of disruption. In 1920, the average length of time of a company in the S&P 500 was 65 years, and the S&P 500 realized only a 1.5% turn-over rate per year at that time. Today the average S&P 500 company lasts only on average 20 years, and by 2020 this turnover rate is expected to hit 10% per year. [Reference: Cleantech Group LLC analysis, from Foster, R. and Kaplan, S., Creative Destruction] So, while disruption of markets is never easily managed, our economy is becoming much more accustomed to rapid change, and it is adapting to this environment successfully. Simply put, business today understands that the best ideas and technology rapidly displace earlier innovations in the markets, and while years ago many companies fought to extend their product life cycles by opposing technological advancement, today there is much more of a philosophy of embracing new ideas, and striving to be on the beneficial side

of these inevitable disruptions. Similarly, universities that boldly innovate today, and which are open to new and much more aggressive strategies of technology management and industrial relations, will in general win big as well. Today, much more so than ever before, the risk of not taking the risk is actually much greater than the risk itself.

**Case Studies**

The fusion of design with technical innovation is essential today. It is not enough to have the best technology, but rather to have the technical designs that are the most readily adapted to the life-styles and work habits of society. In 1997 Michael Dell said that "If I ran Apple, I would shut it down and give the money back to shareholders". At that time Apple was struggling to survive, with rapidly shrinking demand for its products, which were generally considered to be interchangeable with PC-based computing products like Dell computers. Apple had a market cap of just a few hundred million dollars in 1997, when Dell's market cap exceeded $20B.

Then Dell, in my opinion, fell into a common trap: They assumed that the demand for computers was fixed, and that the only way to prosper was to make incremental improvements that only slightly improved their very narrow profit margins. At the same time, in 1997, Apple brought back Steve Jobs as their CEO, and he worked tirelessly to place Apple back on a path of compelling new product development that successfully redefined the market demand for computing and communicating devices. Today Apple's market cap approaches $600B, following a $13B profit in the fourth quarter of FY2011 alone, making it one of the very the top profit quarters of any company in history.

Meanwhile, Dell has demonstrated lackluster performance, with a market cap that remains today about where it was in 1997. The important point here is that innovation and customer-responsive design are critically important, and that the opportunity cost of not taking a risk is often far greater than the possible down-side of the risk itself. Universities, which are historically very risk adverse, often operate more like Dell than like Apple, in that they pass huge opportunities in favor of managing the status quo. This is most unfortunate, since usually a university's market position can be preserved and expanded when innovative risks are explored at low cost, if the process is managed properly.

Apple's successes were not derived from putting the current product line in 1997 at risk, but rather were obtained through the careful development of very innovative new products that were only released to the public when their compelling nature was clearly evident. In my opinion, universities, like Apple, are in an excellent position to pursue remarkable new innovations without taking substantial business risk since they are not heavily invested in the current technologies on the market, except in educational technologies.

It is important to realize that modern aviation emerged out of a bicycle shop in Dayton, Ohio, and most of modern physics, including the development of relativity and the seeds of quantum mechanics, emerged from a low-level

patent clerk in Germany. Too often major universities fall into the trap of thinking that substantial developments must be achieved by following the leadership of the government or other large and more bureaucratic entities. In contrast, history displays time and again that exceptional innovation moves quite disruptively against these well-established trends, without attempting to be disruptive per se. Orville Write said, "If we all worked on the assumption that what is accepted as true is really true, there would be little hope of advance". He said these words at a time when everyone assumed that innovation through carefully controlled experimentation with air foils was futile, and generally a waste of time of 'dreamers' who were incapable more gainful work. In fact, on October 9, 1903, the New York Times reported on the widely accepted opinion of the day when they printed "The flying machine which will really fly might be evolved by the combined and continuous efforts of mathematicians and mechanicians in from one million to ten million years."

Remarkably, on that very same day Orville Wright reported in his diary that "We started assembly today", in reference to the actual flying machine that would successfully sustain the first powered human flight by the end of the calendar year. The point here is that the determined efforts of visionaries have been the only thing that has redefined our thoughts and that have vastly improved our quality of life. So while it is easy for the common wisdom in Universities to think that small visionary efforts cannot create revolutionary advancement, in fact it is the only thing in retrospect that has.

While large projects are easily defined today in established efforts, such as genetics and the development of new biotechnologies, universities must remain open in a decentralized way to support innovators who are thought to be pursuing wild, out-of-the box approaches. There is certainly room for both in our major universities, as long as our leadership does not restrict this diversity through an exclusive demand for only centralized, large, and only interdisciplinary collaborative projects that are managed from the top. Large, interdisciplinary projects certainly have their place in the rapid development of new markets around well-defined new science that is based upon much earlier innovations, but the process of discovery itself is almost always decentralized, undervalued, and it appears from unexpected sources.

Many innovators, such as Bednorz and Mueller who shared the Nobel Prize for high-temperature superconductivity, reported after the fact that they had to conduct their research in a clandestine manner to avoid cancellation of their work by the leadership at IBM, Zurich. Too often large-scale research leaders demand a monotonic approach to research project development, and this stifles genuine innovation that often leads to revolutionary discoveries and epic new opportunities. In short, it remains critically important for university leaders to permit innovative faculty members to continue to do research that their peers often consider a clear waste of time. Again, this research environment

is naturally adaptive to the decentralized structure of universities as long as administrators do not attempt to exert undue centralized control.

Those universities that insist on top-down management in all situations miss the primary advantage of an investigator-led, decentralized research environment, which often proves to be the most profitable aspect of a research university operations in the long run. Of course this does not preclude large, structured research, development, and production operations within universities as well. It just stresses the need for balance, and respect for letting really smart people, typically with tenure, do what they are genuinely inspired to do.

My final case study concerns the development of radio and modern electronics. The early development of the technical innovations by Maxwell, and independently by Hertz, that led to the first transmission of electromagnetic waves in the 1860's remained little more than a novel laboratory curiosity. Then Marconi and Tesla, working separately and often in fierce competition with each other in the late 1890's, developed a string of patents that resulted in early radio devices that realized a small market among typically wealthy customers on ships and in remote, polar regions where no other form of communication was possible. This spurt of innovation was so intense and unusual that it would be much later, in fact in 1946, before the United States Patent Office would overturn Marconi's patents in favor of Tesla's original innovations that led to modern radio.

Still, radio remained an elusive technology, with every radio set being made slowly and at great expense by only highly skilled craftsmen. It wasn't until Edwin Armstrong discovered and patented two important innovations, namely regenerative amplification and heterodyne in the late 1910's that a clear path to the mass production of radios, and later televisions, became clear. These Armstrong patents became the basis that the Radio Corporation of America (RCA) was built upon, and permitted radio to scale to the point where this new modality of broadcasting became available in almost every home within the industrialized world.

The important thing to realize here is that the original discovery of electromagnetic waves, while foundational, was a 'law of nature', and hence not subject to patent protection. The wave of innovations in radio from Marconi and Tesla produced the first patented devices that were based upon this law of nature, but which lacked the detailed systems understanding that would later be provided by Armstrong, which really made this new technology scale. Here, as in many cases, innovations come in waves, with the much larger market penetration come only much later, when the fundamental organizational principles are discovered.

Today universities can take advantage of these natural transitions through interdisciplinary innovation teams that critically evaluate and improve basic discoveries as they emerge, with a focus on how to scale up on the product demand to huge levels. We have such a focused innovation team

that is active at MU, called the Biodesign Program, which is modeled after a program by the same name at Stanford University. In short, this interdisciplinary design team systematically observes surgical applications of existing operating room technology, and they propose design modifications, some incremental and others fairly revolutionary, that substantially advance the surgical processes in ways that surgeons will rapidly adapt once they are available in the marketplace.

This concept is being broadened from medical instrumentation today to include comparative medicine, and veterinary medical applications. MU is a Wallace Coulter Foundation Translational Partner, which provides MU with one million dollars a year for five years to develop often revolutionary new biomedical instrumentation. Many programs and faculty member research groups within MU, including the Biodesign Program, compete for these funds to develop market viable biomedical devices that are based on MU's intellectual property. Programs such as this and others at MU provide the necessary innovation and resources to build upon initial procedures that may be made to scale to achieve much greater levels of market penetration. The investment in this type of systematic later phase innovations on existing products is a generally low-risk approach, with substantial payoffs, as the Coulter Foundation has demonstrated at fifteen different universities across the United States.

It is interesting to note that the advancement of materials science is often the key to the development of earlier innovations to the point where they scale dramatically, and many properly managed materials programs at universities throughout the world have proven their profound value in this regard. As an example, consider the development of the transistor by Bardeen, Shockley, and Brittan at Bell Laboratories in 1948. This discovery required a wave of new germanium and silicon processing technology before it could truly revolutionize microelectronics to the point where it is today. In the 1950's and into the 1960's transistors were made our of poorly processed and controlled materials, so typically post-production sorting, transistor by transistor, was necessary in order to separate the one in 50 or so devices with exceptionally high performance from the more common ones of adequate performance, and from the majority that simply didn't work. It would take the development of a new process for germanium and silicon purification, called zone refining, at Bell Labs in the 1960s before transistors could be made to work well reliably.

In fact, once the materials purification and refining would permit millions and later billions of transistors to be made to work well within very tight tolerances, the opportunity for a transition to an entirely new level (that of large, integrated microelectronic circuits) would completely revolutionize electronics and issue in the modern microelectronics age, which continues to evolve rapidly according to Moore's Law today. All these are examples of predictable waves of innovation that build upon an initial discovery to create a scalable, new industry of profoundly large propor-

tions. Universities that prosper will invest intelligently in such teams, such as the Biodesign Program and advanced Materials Programs, to take full advantage of the later waves of innovations that are based predictably on major new discoveries as they emerge.

**Acknowledgements**

The author acknowledges many useful discussions with Dr. Paul Dale and his colleagues of the MU Biodesign Program, Dr. Jake Halliday and his colleagues at the Missouri Innovation Center (MIC) and the MU Biosciences Incubator at Monsanto Place, and with Dr. Tom Skalak, Vice Chancellor for Research at the University of Virginia, and the Principal Investigator within the University of Virginia's very successful Coulter Foundation's Translational Partnership.

# Developing Infrastructure for Informatics Research: Experiences and Challenges

**Prem Paul,** Vice Chancellor for Research and Economic Development
University of Nebraska-Lincoln

Scientific advances are generating massive amounts of data, fostering the need for new strategies to access, store, analyze, and mine data[1] [2]. The need to manage "big data" is especially acute in the life sciences, where genomes of a large number of species have been completed and efforts are underway to correlate genetic information with biological functions. Efforts are also underway to identify genes associated with health and disease. Similarly, large international collaborative experiments in physics, such as those conducted at CERN's Large Hadron Collider that recently resulted in the discovery of the Higgs boson particle, are generating large amounts of data and requiring high speed connectivity between laboratories to transfer data and to support high capacity data storage and analysis.

Most institutions are trying to deal with these challenges, which require major financial investments in infrastructure and personnel, resulting in significant economic pressures at a time when most institutions are facing budget cuts and federal funding is expected to flatten or be reduced. At the University of Nebraska-Lincoln (UNL), we recognized early on the need for enhanced cyberinfrastructure to support our researchers, and we initiated discussions on this important topic in 2005. We held an all-university workshop attended by 150 faculty members from the life and physical sciences, engineering, and the humanities. This paper summarizes our experiences, challenges, and plans for dealing with big data.

**Advances in Life Sciences**

Advances in nucleic acid sequencing technology have made it possible to sequence complete genomes of a large number of species. Information on thousands of genomes is now deposited in the National Center for Biotechnology database (see Table 1 for a summary of some of the major genomes).

This information makes it possible to determine biological functions coded by various genes and also to determine the significance of various genes relevant to health and disease. Nucleotide sequence data is being utilized to identify genes associated with cancer and other diseases and to develop novel therapies. At UNL, our faculty are working on plant and animal genetics and utilizing genomics in their research to improve productivity, particularly regarding traits for disease resistance and/or drought tolerance. These studies require major investments in computing and bioinformatics infrastructure and personnel trained in bioinformatics.

**Cyberinfrastructure Needs in High Energy Physics**

In 2005, UNL faculty identified the need for enhanced infrastructure for computing and connectivity. Our faculty competed for a National Science Foundation-funded Compact Muon Solenoid (CMS) Tier-2 site in high energy physics that utilizes data generated by CERN's Large Hadron Collider near Geneva, Switzerland, for research through the Department of Energy Fermilab. This research project required at least 10Gb Dark Fiber for data transfer and enhancement of computing infrastructure. We held a cyberinfrastructure workshop with experts from various funding agencies, including the DOE and National Science Foundation (NSF) and other institutions to learn about the importance and current state of cyberinfrastructure more broadly. NSF had published a blue ribbon cyberinfrastructure report[3] that was used as a background material. As a result of the workshop, we decided to invest in 10Gb Dark Fiber connecting Lincoln and Kansas City at a cost of over $1 million to connect with Internet2.

This gave our faculty the ability to be in a leadership position and transfer a record amount of data from Fermilab. UNL's leading capability was demonstrated at a national Internet2 meeting in 2007. The cyberinfrastructure we developed to manage big data has also been very helpful in supporting our faculty who require supercomputers. For example, leveraging these computing resources, Professor Xiao Cheng Zeng in UNL's Department of Chemistry has made several major discoveries published in top-tier journals like the *Proceedings of the National Academy of Sciences* [4 5].

UNL's Center for Digital Research in the Humanities has been a leader in the digitization of scholarly material related to Walt Whitman and the Civil War. The group of faculty in this center have so successfully competed for grants from the National Endowment for Humanities that they are recognized as national leaders in this area.

**Bioinformatics Experience**

One area in which we have made significant investments is bioinformatics. We have hired several faculty members during the last decade with bioinformatics expertise. They have been very successful scholars and have been extramurally funded with grants from the DOE, National Institutes of Health, NSF, and U.S. Department of Agriculture. However, they are pursuing their own scholarship and research agenda and are not able to provide bioinformatics service to others. A large number of faculty in the life sciences who do not have a background in bioinformatics

| Species | Genome Size | Predicted Genes Coded |
|---|---|---|
| Arabidopsis | 119 Mb | 25,000 to 31,000 |
| Fruit Fly | 165 Mb | 13,600 |
| Mosquito | 278 Mb | 13,700 |
| Rice | 420 Mb | 32,000 |
| Corn | 2300 Mb | 32,000 |
| Mouse | 2500 Mb | 23,000 |
| Cow | 3000 Mb | 22,000 |
| Human | 3400 Mb | 20,000 to 25,000 |

Table 1. Summary of major sequenced genomes

need help with the analysis of sequence data, and are having a difficult time finding experts to do their work. There also is a shortage of talented people trained in bioinformatics.

We have a bioinformatics service in our Center for Biotechnology core facility; however, the facility's staffing is not sufficient to meet the needs of all faculty because there are more users than talented experts. In our experience, life scientists want bioinformatics experts to analyze their data – much in the same way statisticians have contributed to research programs for decades. However, the majority of the bioinformatics experts want to pursue their scholarship and advance the bioinformatics field. Several research groups have created their own bioinformatics core facility, including their own computer clusters, rather than using supercomputers. We are exploring ways to add bioinformatics staff in the core facility and hire additional bioinformatics faculty, including a leader who can coordinate bioinformatics resources and services across campus.

**Big Data Needs in the Social and Behavioral Sciences**

There are also needs for access to big data and cyberinfrastructure to address important questions in the social and behavioral sciences. UNL has significant strengths in social and behavioral sciences, including the Gallup Research Center, which conducts research and trains graduate students in survey methodology. The Bureau of Business Research provides relevant information and insightful data on economic conditions across Nebraska, the Great Plains, and the nation. UNL's Bureau of Sociological Research and our Survey, Statistics, and Psychometrics Core Facility provide services to faculty in survey methodology and research. The University of Nebraska Public Policy Center provides the opportunity for policy makers and researchers to work together to address the challenges of local, state, and federal policy.

We also have strong programs in substance abuse and health disparities in minority populations. Though each program is highly successful, there is an opportunity for strengthening these programs through collaborations. This is critical, especially considering the importance of social and behavioral sciences in major societal challenges pertaining to food security, water security, national security, economic security, national competitiveness, and energy security. Therefore, we have launched a taskforce to better understand our institutional strengths and needs for infrastructure.

UNL faculty members have recognized the need for a Research Data Center (RDC) to access economic, demographic, health statistics, and census data. RDCs are run through the Census Bureau and NSF. Currently, there are 14 RDCs managed by the Census Bureau. RDCs provide secure access to restricted use of microdata for statistical purposes. Qualified researchers prepare proposals for approvals by the Census Bureau. Following approval, work is conducted in secure facilities where scientists may access centrally held data.

Current RDC locations include: Ann Arbor, MI; Atlanta, GA; Boston, MA; Berkeley, CA; Chicago, IL; College Station, TX; Ithaca, NY; Raleigh, NC; Stan-

ford, CA; Washington, D.C.; Minneapolis, MN; New York, NY; and Seattle, WA. Unfortunately, there are no RDCs in the Midwest; the center nearest to Nebraska is in Minnesota.

Since RDCs are expensive to maintain and require hiring a director that is a Census Bureau employee, it might be more appropriate to pursue a regional RDC that could serve universities in Nebraska, Iowa, Kansas, and Missouri. Based on conversations with Census Bureau personnel, such an RDC would comprise secure space, including workstations for faculty and students to access data for research. We propose to build such a center at UNL that would be available to our regional partners. Access will be facilitated through proposals that are peer-reviewed by an advisory board, as required by the Census Bureau protocols.

Several years ago at the Merrill Conference, discussion took place regarding what we can do together that we cannot do alone – especially with regard to creating shared research infrastructure to support large-scale research projects and programs. The RDC concept represents such an idea for regional collaboration to access big data in social and behavioral science research.

**References Cited**


1. T. Hey, S. Tansley, and K. Tolle (Eds). 2009. *The Fourth Paradigm: Date-Intensive Scientific Discovery*. Microsoft Research: Redmond, Washington.
2. P.C. Zikopoulos , C. Eaton, D. deRoos, T. Deutsch, and G. Lapis. 2012. *Understanding Big Data: Analytics for Enterprise Class, Hadoop and Streaming Data*. McGraw Hill.
3. D.E. Atkins, K.K. Droegemeier, S.I. Feldman, H. Garcia-Molina, M.L. Klein, D.G. Messerschmitt, P. Messina, J.P. Ostriker, and M.H. Wright. 2003. "Final report of the NSF Blue Ribbon Advisory Panel on Cyberinfrastructure: Revolutionizing Science and Engineering through Cyberinfrastructure." Available at www.nsf.gov/od/oci/reports/toc.jsp.
4. J. Bai, C.A. Angell, and X.C. Zeng. 2010. "Guest-free monolayer clathrate: coexistence and phase transition between two-dimensional low-density and high-density ice." *Proceedings of the National Academy of Sciences USA*: 107, 5718-5722.
5. T. Koishi, K. Yasuoka, S. Fujikawa, T. Ebisuzaki, and X.C. Zeng. 2009. "Coexistence and Transition between Cassie and Wenzel State on Pillared Hydrophobic Surface." *Proceedings of the National Academy of Sciences USA*: 106, 8435-8440.

# Skating to Where the Puck is Going to Be

**Steven Warren,** Vice Chancellor for Research and Graduate Studies, University of Kansas

"A good hockey player plays to where the puck is. A great hockey player plays to where the puck is going to be." So goes a famous and often quoted observation attributed to the great professional hockey player Wayne Gretsky. Whether this is actually a good recommendation for a hockey player is beside the point. The statement captures the challenge that is especially pertinent for the legions of professionals trying to figure out where the ongoing revolution in informational technologies is going so that they can "be there when it arrives" instead of lagging behind.

The purpose of this essay is to offer my perspective, as a Chief Research Officer at a Midwestern university, on a few of the challenges we face as we skate forward into the future. My essay is divided into three sections. First, what is it that we really want from information technology? Second, what are a few of the big picture issues generated by the IT revolution and their local relevance? Finally, I offer a few closing thoughts.

**What do we want?** If you are on the business side of the IT revolution, you want to know: "What does the consumer want?" That's where the consumer market is and that is what big businesses (e.g. Microsoft, Apple) are going to try to satisfy. There are lots of things we'd all like to have. These include transparency, convergence of technologies, simplicity (meaning that we want the complexity to be hidden by straightforward and intuitive interfaces), and of course we want safety and security from hackers, thieves, etc. Yet as important as these needs are, two even more basic elements top them all. These are **SPEED** and **POWER.** An obvious example of speed and power at a basic consumer level is the progression from the first Iphone to Iphone5. The first Iphone was amazing. But the Iphone5 (as was the case with Iphone 2, 3, and 4) beats it in terms of processing speed and the wide range of things it can do with this speed – that is its power.

Now translate the same concepts into what scientists want. First they want to assume all the same basic stuff - meaning they want convergence, transparency, simplicity and safety/security. But what they REALLY crave is speed and power. In our case (speaking with my scientist hat on now), we want the speed and power to communicate, teach, and learn from anywhere in the world at any time, with ease. We want to have the power and speed to analyze remarkably complex problems. These include the ability to study the most complex known object in the universe (the human brain), the basis of life in all its forms, our planet and how it operates and behaves, and of course the universe itself.

Yes, we are indeed an ambitious and pretentious species - one with an almost unquenchable thirst for SPEED and POWER. Furthermore we have become totally spoiled over the past 40 years as a result of Moore's law and the remarkable skills of a small group of computer scientists and engineers who exploited this so called law. Moore's law observes: "…over the history of computing hardware, the number of transistors on integrated circuits doubles approximately every two years" (Wikipedia). What this doubling has enabled is the remarkable virtually exponential gains in SPEED and POWER every two years. Exponential growth is so fast that most of us struggle to even conceive of what it actually means, let alone how to keep up with it. But it has enabled extraordinary breakthroughs in science, education, entertainment, transportation, and on and on. It is literally transforming our world in countless ways.

However, we may be nearing the end of this incredible ride. Various individuals and groups predict that we are very close to the end of Moore's law, or at least to it slowing down. Some even think that the end of Moore's law will have huge negative impact on economic development with devastating consequences. Others believe it will be little more than a speed bump in the road and that breakthroughs in other areas (e.g. nanotechnology) will keep pushing us rapidly ahead. Actually, if things did slow down a bit, that will have its positive effects too – such as allowing as to consolidate all of our technological breakthroughs and catch up just a bit.

**Skating to Where the Puck is Going in an Era of Radical Change**

Many of the real impacts of the remarkable changes in information technologies are just emerging on the horizon. These changes are already having a big impact on the higher education enterprise in general and a tsunami of disruptive change appears to be roaring right at us. So how does one manage a big, complicated university research enterprise in this environment? What are some of the changes that are already roaring through the Ivy Tower? There are lots of examples of these. I'll share just four unique examples of the changes underway in the world of research.

1. ***The infrastructure of research administration.*** One of my goals is to transform the research administration experience for scholars at the University of Kansas by creating a fully integrated electronic research administration system. Most of our system is already electronic of course. But it is not integrated in any way that allows people to efficiently and effectively manage it. I want to put the Investigator at the head of the line – they are the reason for research administration, they are the customers. They are the ones that need a straightforward easy way to use systems that will make the administration of their research easier, instead of more complicated.

   Here is what that might look like: I, Mr. Researcher, flip open my laptop anyplace in the world, put in my password, and open my **personal faculty research portal**. Everything I need to effectively manage my

grants and projects is right there. I can check the financial balance of my grants, look at projections given my present rate of spending, update and submit a request to the Institutional Review Board, work on a new proposal, submit and monitor a travel request, and on and on. This sounds reasonable enough - we just need to get these systems to converge and all start talking the right language.

In fact, it's been a big challenge. We haven't solved it yet, but we are on the path. When we started down this path several years ago, it became obvious to me that we needed someone really comfortable in this integrated electronic world. And so, I now have a 28 year old Assistant Vice Chancellor overseeing our entire research administration system. He is capable of leading this change while those with much deeper research administration experience may struggle because they lack his technological sophistication and comfort.

2. ***Avoid the front of the line.*** In the rapidly changing world we live in, the most exciting place to be is of course on the front end of innovation. It can also easily be the most expensive, complicated, and disappointing place to be. Why? First, often at the front of the line you pay the highest price for something because market forces have not yet taken over (note: it can sometimes be cheaper too because companies are selling low in order to break into a market). Second, lots of things don't work very well right out of the box. Their performance improves with time and experience, because companies know they must make those improvements or ultimately the customers will walk away. Third, sometimes new innovations simply fail when they go to a larger scale. We had this exact experience at KU when we signed up with a company that was in the process of creating a "PI portal" just like I described above. It didn't end up costing us very much money, but it did cost us a lot of time. Ultimately we will achieve our goal, but the wasted time and effort associated with being an early adopter was a sobering experience. Bottom line, letting others serve as the early adopters may mean that you get a better, more reliable and cheaper product in the end.
3. ***The changing nature of research collaborations.*** In my experience, the hyper competitive world of research in combination with the hyper connected world of we live in, has resulted in profound changes in collaboration amongst scholars. It is still the case that we like to collaborate with colleagues who work near us - all other things being equal. This kind of collaboration can be relatively easy and are sometimes especially creative. But we generally don't collaborate with people just because we like them or they are nearby. Instead we most often collaborate because we need to in order to be successful, and often because it's the only way to bring a sufficiently wide range of

skills to bear on some complicated research problem we face.

A fundamental outcome of the IT revolution has been the change it has made in terms of how easy it is to collaborate with anyone almost anywhere. We easily collaborate with people all over the world…people that we may rarely see (expect perhaps on computer screen). We may still want to have some actual physical contact with these collaborators to build a sufficient level of trust. But that's it. Our technologies allow us to solve virtually all of the analytic and technical problems that in the past would have stymied these types of collaborative efforts, or simply made them too expensive and complicated to do. As a result, my experience has been that scientists will seek out whoever they need to solve the challenges they face, often without regard to location. This has contributed to the explosion of scientific knowledge over the past couple of decades.

4. ***Scientific Fraud is becoming much easier to catch.*** With all the reports of scientific fraud over the past decade, it would be easy to assume that many scientists have lost their moral compass and are trying to cheat their way to fame and fortune. It is true that the pressure to be successful, to stay funded, may have increased and thus contributed to the increase in fraudulent data. But maybe it was there all along and was just too hard to catch. We will probably never know, but we do know that it is getting much easier to catch certain types of scientific misconduct due to breakthrough technologies that themselves are just another side effect of the IT revolution.

The biggest change is in our ability to detect plagiarism. Scientific journals can now subscribe to services that will scan each submission they receive and compare it to countless related papers that have been published all over the world, in a search to detect instances of plagiarism. Some people even make this a kind of hobby – finding papers in the literature that contain a significant amount of plagiarized material and then reporting the alleged perpetrator. Because of this, I anticipate that plagiarism will all but disappear as a scientific concern in the near future.

And the most serious type of scientific misconduct – publishing false or fabricated data – may not be far behind, as powerful techniques are perfected that can detect highly unlikely findings that at the minimum, need external replication to determine their accuracy and validity. These kinds of innovative tools have recently been applied to some areas of social science where the crucial tool of independent replication has rarely been used in the past to identify questionable findings that need to be further tested. Will scientific fraud eventually vanish in our hyper-connected world? Probably not completely, but the quality and reliability of science overall is already being significantly improved by breakthrough tools that owe their existence to the IT revolution.

**Final Thoughts**

We live in truly remarkable times. The pace of technological and scientific innovation is staggering. The foundation for much of this is the ongoing tsunami generated by the information technology revolution. This tsunami is washing over higher education and may ultimately radically transform many aspects of what stands as perhaps the most enduring institution of the past five hundred years. Universities in fact are the origin of much of this disruptive, creative destruction that is rolling across virtually every corner of the world. But being part of the source of this revolution does not in any way inoculate us from its transformative effects. Consequently some of the really big questions remain to be answered. For example, will the basic model of research universities survive the exponential changes in information technologies? That and so much more remains to be determined. In the meantime, Wayne Gretsky's famous quote remains mostly an aspirational goal. The puck is speeding ahead of us exponentially, and spreading out in many different directions. Nevertheless, we must keep trying to skate to where we think it is going.

# Information as a Paradigm

**Jeffrey Scott Vitter,** Provost, University of Kansas

## Introduction

Kansas has always held a unique distinction as the "center" of the United States, as determined by the technology of the day. Using the best technologies available in 1912 — the year New Mexico and Arizona became states and completed the Lower 48 — the U.S. National Geodetic Survey determined the nation's geographic center to be near Lebanon, Kansas. Of course, technology has changed dramatically since then, and today advanced mathematical models say that the geographic center is actually 20 miles west in Kansas, closer to Phillipsburg.

Similarly, the presumed location of the nation's geodetic center, which is different from its geographic center, has also changed over the years. In 1927, a point in Kansas known as Meades Ranch, about 40 miles south of Lebanon near Lucas, was declared the geodetic center of North America. It held this title for the next 56 years until new technologies led to the establishment of the North American Datum of 1983 and the World Geodetic System of 1984.

Today, the most modern mapping technology is Google Earth, earth.google.com. But guess what? Google Earth *still* lists Kansas as its default center. If you zoom in when Google Earth opens, you arrive at an aerial view of Meadowbrook Apartments, right across the street from KU! Admittedly, that fact has less to do with mathematical models and more to do with the fact that Google Earth creator Brian McClendon is a proud University of Kansas graduate whose world once revolved around KU. And his apartment in Meadowbrook is now memorialized each time someone opens Google Earth.

The point is, information and technology — and their intersection in the area of information technology (IT) — matter a lot and continue to change the world. Information technology is the breakthrough development that has opened all kinds of doors for society and civilization. This paper proposes information technology as a paradigm, both for advancing our agenda at KU in research excellence as well as for a basis of everything we do.

**Information Technology as a Paradigm**

We are in an information age. Computer science, information, and IT have made huge advances in the last few decades. Now is the time and place for them to have a major effect. They are powerful tools, ready to be used in all sectors of society. Advances in computer technology have fundamentally changed the way we live. In fact, computer technology has become the infrastructure that drives commerce, entertainment,

healthcare, national security, transportation and innovation.

In 2009, a panel of eight judges from the Wharton School of the University of Pennsylvania named the top innovations of the past 30 years. The panel received some 1,200 suggestions, ranging from lithium-ion batteries, LCD screens and eBay to the mute button, GPS and suitcase wheels — a list that illustrates the incredible pace of innovation over the past 30 years. Not surprisingly, most of the top 20 innovations on the list were technological and medical advances. Also not surprisingly, the Internet topped the list. And nearly all of the top 20 innovations were directly tied to IT or influenced by IT, specifically computer science.

For the first few decades of computer use, IT did not have a measurable effect upon the economy and economic productivity. But starting in the mid-1990s, there has been a dramatic increase in economic productivity in the U.S., and IT is the major driver. Harvard economist Dale Jorgenson and colleagues determined that, in the years following 1995, IT accounted for the majority of the growth in productivity in the U.S. economy[1].

The National Research Council's Computer Science and Telecommunications Board illustrates the growth of the IT economy with a chart commonly referred to as the "tire tracks" diagram (see Figure 1).[2] The lines of the tire tracks diagram resemble the grooves left by a tire — the thin red line on top indicating when research was performed in universities, a thicker blue line in the middle shows when research labs were working in the space, and a dotted black line indicates products being introduced. When the dotted black line turns green, that indicates when the technology became a billion-dollar sector.

When presented this way, there are three very clear takeaways that emerge about the growth of these billion-dollar sectors: First, each one of these sectors can trace its formation to university research and, in almost all cases, to Federally-funded research. Second, it takes a long time for the research to pay off, in most cases one or two decades. And third, the research ecosystem is fueled by the flow of people and ideas back and forth between university and industry. This system has made the United States the world leader in information technology.

**Technology as the Infrastructure for Grand Solutions**

KU has recently embarked upon a transformative strategic plan to chart the university's path toward excellence as a top-tier public international research university. Aptly titled *Bold Aspirations*, boldaspirations.ku.edu, the plan sets out the new, higher expectations KU has for itself as a university and the priorities that we will pursue. As part of the plan, KU research is being targeted toward four societal "grand challenges"— challenges so complex that they cannot be solved by any one discipline and require inherently multidisciplinary approaches. We call these four grand challenges our strategic initiative themes:

"Sustaining the Planet, Powering the World" — focusing upon energy, climate, and sustainability

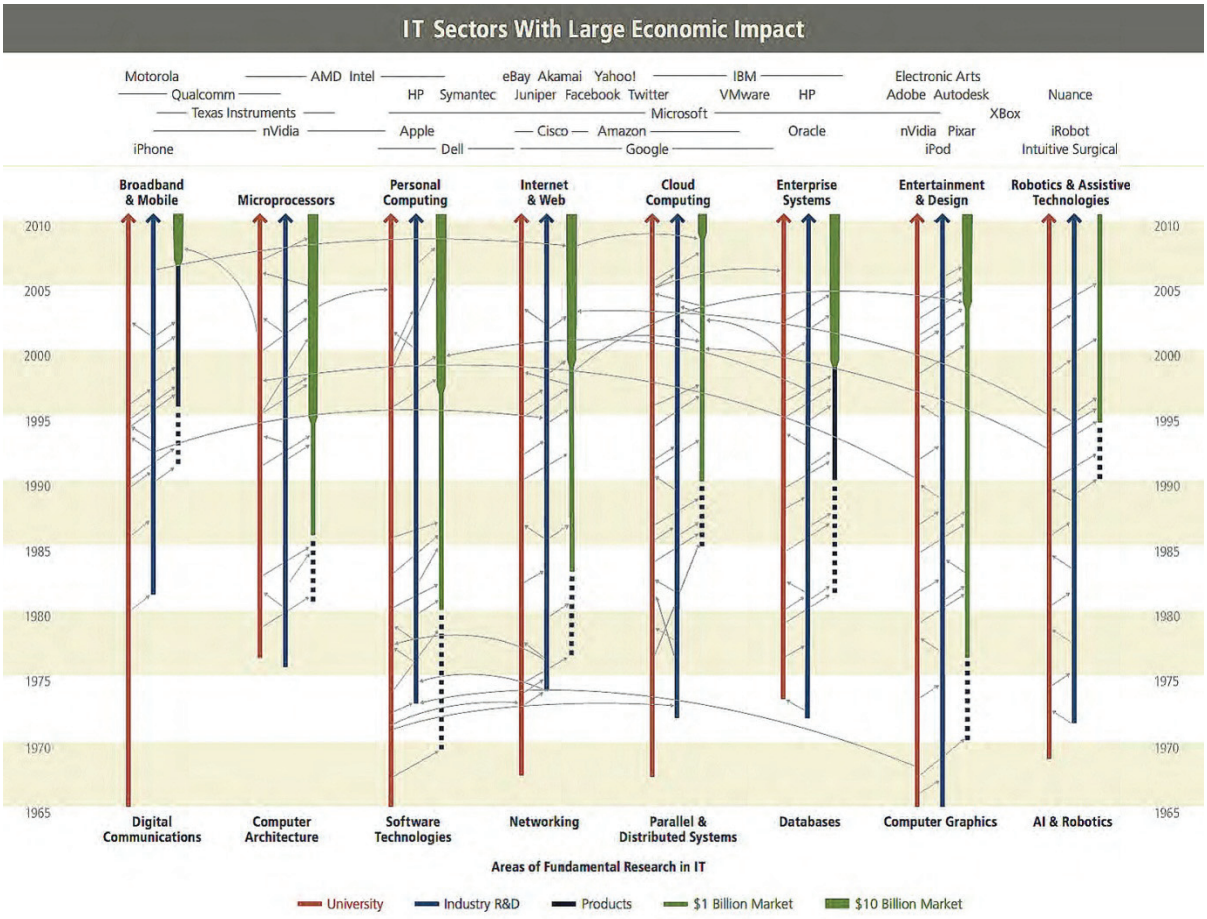



**Figure 1:** The "tire tracks" diagram, so named because of its visual resemblance to the tracks that an automobile tire makes, was put together by the National Research Council to represent the tremendous growth of various sectors of the IT economy over the last few decades. Each sector is represented chronologically by a vertical column, with time moving from bottom to top. The red line on the left for a sector indicates when Federally-sponsored research was performed in universities. The middle blue line indicates corporate research in the sector. The dotted black line indicates when the first product was introduced in the sector, and the time at which the dotted black line turns solid green indicates when the sector became a billion-dollar industry. Where the line thickens indicates a $10 billion sector. The gray arrows indicate the flow of people and ideas among sectors. Some of the specific billion-dollar companies that have emerged from these developments are listed at the top of the chart.

- "Promoting Well-Being, Finding Cures" — focusing upon health and well-being
- "Building Communities, Expanding Opportunities" — focusing upon local, national, and global communities
- "Harnessing Information, Multiplying Knowledge" — focusing upon the transformative power of information

It's the fourth theme listed above — "Harnessing Information, Multiplying Knowledge" — that forms the core of this paper. It is crucial components of the previous three themes and drives the future of KU. IT is, in fact, a paradigm for what the university does.

A great example of the role of IT across research areas can be seen in the list of 14 grand challenges in engineering assembled by a committee convened by the National Academy of Engineering, listed at www.engineeringchallenges.org. According to some observers, notably Ed Lazowska at the University of Washington, eight of the 14 goals on the list require a predominant role in computer science. In the six remaining areas, IT will play a significant supporting role by providing the infrastructure for solutions.

**Using Technology as Infrastructure at KU**

IT is not only a key driver for research at KU. We are also using IT in a broad sense to build an infrastructure for innovation. One example is our new Center for Online and Distance Learning (CODL), which helps faculty build online or hybrid courses and also serves as a central point for students to access online learning. More specifically, the CODL is using technology to create deeper and better learning experiences, more efficient classroom experiences, and more opportunities for faculty and students.

Another example of how KU is using technology as infrastructure is the Open Learning Initiative, which offers online courses to anyone who wants to learn or teach. KU is also exploring other such hybrid teaching models, which can offer various advantages. For example, online learning modules track student mastery of basic core concepts for a course. This allows the faculty member, through the gathering of data in the online course, to know where and how each student is progressing. Additionally, these models permit students to repeat parts of the course until they attain mastery, allow class time to instead be spent integrating basic knowledge, and can demonstrate a baseline of mastery in all foundation courses in a discipline.

Another example of how KU is using technology as infrastructure is the university's 2009 adoption of an open access policy, which makes faculty members' scholarly journal articles available online for free. KU was the first public university to adopt such a policy. The main points of the policy are straightforward: faculty members grant a non-exclusive license to the university to share a copy of their paper; faculty members give a copy of papers to the university for open dissemination; and faculty members may notify the university of their waiving of the license granted at their discretion for any individual paper. KU Libraries hosts a public portal called KU ScholarWorks that provides

free access to all faculty publications under the policy.

KU is also using technology to maintain Faculty Professional Records Online (PRO), a faculty activity database that serves as a repository for faculty scholarship, simplifies evaluations, creates searchable experts lists, and identifies clusters of strength across disciplines. The PRO database feeds additional outreach, allowing us to highlight our faculty's accomplishments and connect them with various groups such as policymakers, media, entrepreneurs, and other stakeholders. PRO also helps grow our open access portal KU ScholarWorks mentioned above by feeding articles and books by faculty members into the repository.

Last, but certainly not least, we are using technology as infrastructure at KU by building capabilities for sophisticated analytics that allow us to examine our effectiveness and productivity as a university. A good example is our partnership with Academic Analytics. Figure 2 illustrates one of the graphics we generate to study program effectiveness. In the chart, the scholarly productivity of seven different KU programs and department are represented by the red line, contrasted with AAU member schools (in blue) and large research schools (in green). This diagram is useful in a number of ways — like helping to determine which programs to invest in. If we see strong programs (like departments A and B, for example), we can dig deeper to further examine the sources of their strength. Are there pending retirements that will weaken these programs? Would additional senior hires improve a department's national rank to the next level? Can pending retirements and replacements change a department's productivity level? For other cases, we could ask if this program is one that has slipped so far down that we should stop investing in it? Is it a program we may want to transition to "non-research" focus? These are the types of crucial questions and strategies that can be addressed as a result of information technology.

**Culture of Scholarly Engagement**

The state of Kansas demands a flagship university in the top tier of public international research universities. This is a reasonable demand, because as Richard Florida observes, great universities are magnets for innovative minds, and innovation is the key to a flourishing and diversified economy. For that reason, one of the goals of KU's strategic plan *Bold Aspirations* is to promote a vibrant culture of scholarly engagement. To do so, KU continues to actively engage with communities throughout Kansas and the world, with a focus upon entrepreneurship, commercialization of technology, and vibrant business partnerships. All of these depend upon IT.

A great example of this engagement has been the KU Cancer Center's multi-year drive to achieve National Cancer Institute (NCI) Cancer Center designation. In July 2012, KU met this goal. The university and its partners continue to invest in the KU Cancer Center, and university officials plan to apply for Comprehensive Cancer Center status in 2015. The research and investment revolving around NCI designation touches upon each of KU's strategic initiative themes, and IT plays a key role.

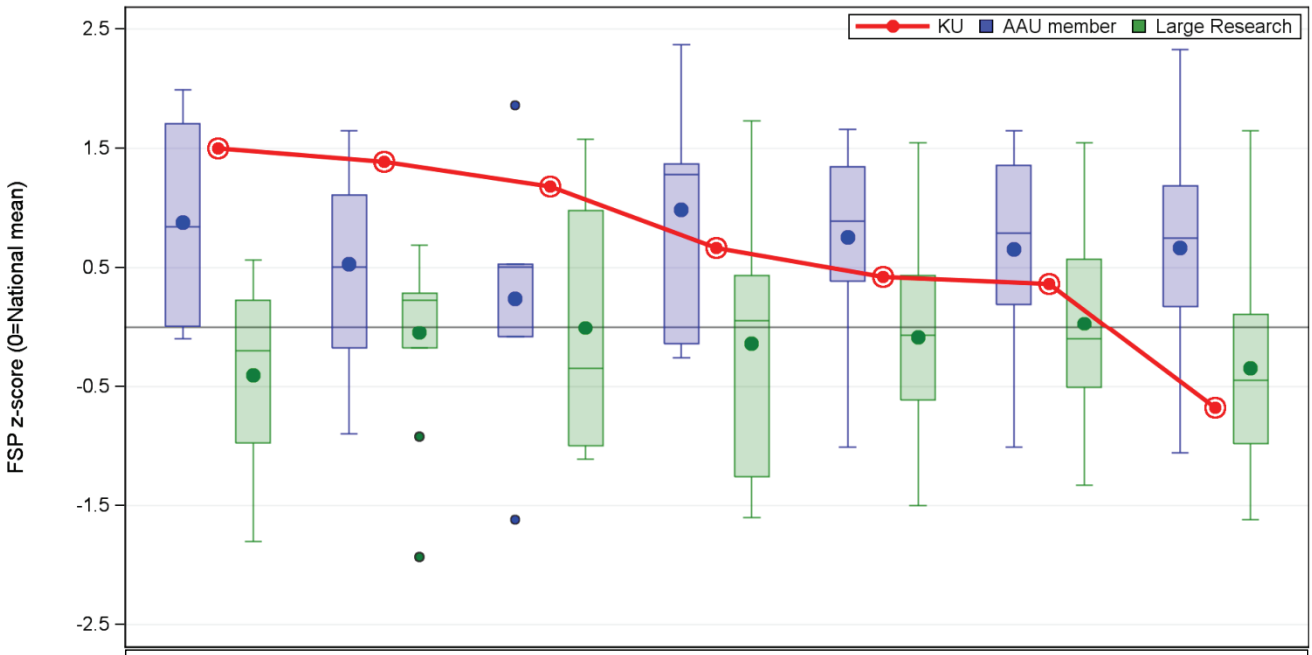



**Figure 2:** This "whiskers" diagram was developed at the University of Kansas to indicate the standing of departments at KU relative to national research university peers using data collected by Academic Analytics. For each of seven academic programs listed from left to right, the red line shows the KU ranking in that program relative to national norms. Each unit represents one standard deviation above or below the mean. The blue box indicates the performance range for the 34 public U.S. universities that are members of the prestigious Association of American Universities, which KU has belonged to since 1909. The box captures the 25th–75th percentile range, with the median indicated by the horizontal bar and the mean indicated by the blue dot. The full range is indicated by the "whiskers" above and below the box, and if there are extreme outliers, those are indicated separately by blue dots above or below. The green boxes similarly indicate the performance of the non-AAU large public research universities.

KU has also recently transformed the KU Center for Technology Commercialization (KUCTC), the entity charged with spearheading technology transfer and commercialization efforts. Over the past two years, KU has hired four national experts to propel KUCTC forward. The first two have built records of accomplishment at Purdue University: Julie Goonewardene, president of KUCTC, focuses upon commercialization, licensing, and startups, and she also serves on the board of the American Medical Association; Julie Nagel directs business relationships and our Strategic Partners Program. Becky Voorheis, an experienced Silicon Valley entrepreneur, assists our faculty with company startups. And Rajiv Kulkarni, our director of technology transfer, came to KU in February 2012 from the University of Utah, a national leader in technology transfer. These hires represent a growing focus upon commercialization and entrepreneurship at KU.

KU is also home to the Bioscience and Technology Business Center (BTBC), a statewide incubator network that has 24 tenants spread across four buildings in Lawrence and Kansas City. The tenants include a diverse range of bioscience and technology-based businesses, ranging from KU startups and growth companies to large corporations like Garmin and Archer Daniels Midland. By locating in the BTBC, tenants get access to KU facilities and researchers and also receive various business support services from BTBC staff. The BTBC system's flagship facility in Lawrence reached 100 percent occupancy just 18 months after opening, and plans are in place to more than double that facility's square footage.

KU will soon launch a faculty expertise portal designed to help users quickly find out who is doing what at KU. The portal will gather data and do semantic matching automatically from our PRO database. Thus, it's a great tool for companies to find faculty with desired expertise. It's also great for potential graduate students to find out about areas of interest, or even for faculty at KU wanting to partner with other faculty on new projects.

KU will also use information technology to support its new Strategic Partners Program, which is designed to identify and nurture partnerships between KU and the business world. We have started use of constituent relationship management software to enhance our ability to interact with prospective students and current students. The same software can be used to track KU connections with industry. It will identify the range of partnerships already in place and, through the use of data analytics, identify companies to expand engagement.

Additionally, KU continues to employ social media and related online tools to "take geography out of the equation." For example, technology is being used to build maps of capabilities and expertise, group stakeholders around topics via social media, and create a virtual presence that lets stakeholders connect with KU from anywhere in the world.

**A Global Paradigm**

While significant 100 years ago, it is no longer so relevant that Kansas is at

the geographic and geodesic center of the continental United States. Much of the entrepreneurial action these days takes place on the east or west coasts. However, it is significant that through IT, we can truly immerse ourselves anywhere in the world, link together key partners, and form vibrant collaborations. IT drives society, and it drives KU. Maybe it is fitting after all — and not just Jayhawk loyalty — that Google Earth is centered on Lawrence, Kansas. Through our research at KU and as a core part of how we operate, we use information in fundamental ways to improve our understanding of the world and to make it a better place.

References


1. Jorgenson, D. J., Ho, M. S., & Stiroh, K. J. (2005). Productivity, volume 3: Information Technology and the American Growth Resurgence. Cambridge, MA: MIT Press.
2. National Research Council (2012). Continuing Innovation in Information Technology. Washington, DC: The National Academies Press.
3. Florida, R. (2003). The Rise of the Creative Class. Basic Books.

# Scholarly Communication in the Age of New Media

**Brian Foster,** Provost, University of Missouri

## The Broad Perspective

It is important to put scholarly communication in context by addressing the very broad idea of why scholarly communication is important. One of the main missions of universities is to create new knowledge and innovations—i.e., research. This new knowledge needs to be disseminated—to be exchanged such that the scholars creating the new knowledge are in touch with others, all contributing to the original ideas and new insights of each other. But scholarly communication is not just about people working at the same time on problems that are somehow connected; it is also about archiving research results for investigators of the future. In addition, it is critical that higher education is very much about educating students to be creative and innovative and, in this sense, access to research results and processes is critical for effective education. In short, scholarly communication is critical for both the research and educational missions of universities.

Given the centrality of scholarly communication to the mission of higher education, it is unsettling that all we know about the current model is that it will not work in the future. There are many reasons that this is true, many related to new media technologies that have dramatically disrupted the business models for both print and digital scholarly publishing—i.e., scholarly journals, university presses, and other organizations that publish research results. This is not a new idea: it was spelled out as far back as 1980 in the White House Conference on Libraries and Information Services (Lamm, 1996, p. 127). Journals are pretty much going digital. Books are moving rapidly in that direction, and the future of university presses as we know them is highly problematic (Withey, et al., 2011). It is important to emphasize that scholarly publishing is just one part of scholarly communication—which is part of the problem, as will be discussed in detail below.

Journals and books fit into a broad scholarly communication environment in rather different ways—into the broad mix of "media forms" that include both formal and informal transfer of information. For example:

Informal communication among scholars is in some ways the most important kind of scholarly communication. People working on similar problems often discuss their work informally at meetings, over a meal, having a beer—completely unstructured conversations that are timely as the work is being done. In addition, people working on different kinds of problems may discover entirely unexpected connections

between their areas, opening exciting new directions for their research.

Oral presentation of research results occurs in many different kinds of venues, generally when the work is far enough advanced that at least preliminary results can be reported. Such presentations occur in formal meetings of, say, scholarly organizations, in symposia, or as guest speakers in many other kinds of venues. This tends to be relatively early in the history of a project.

Often there is pre-publication distribution of manuscripts to colleagues working in similar/connected areas.

Proceedings from scholarly meetings are often the earliest and most critical versions of articles in journals—generally significantly earlier than publication in classic, high-ranking scholarly journals. Such proceedings may often also be a significant archiving function—especially for "papers" that are not subsequently published in "journals"—preserving the work and making it available to future generations of scholars.

Articles in scholarly journals are the gold-medal scholarly communication media in many academic disciplines. This kind of publishing is critical for faculty to be promoted and otherwise recognized, and it is a critical archiving function that makes scholarly results available for generations of scholars in the future. It comes late in the research project, however—generally after the scholarship is complete, and is often of little immediate relevance to active research.

Edited books of articles and chapters provide another medium for disseminating scholarly outcomes; as with journals, the importance of such publications varies by scholarly discipline.

Monographs provide the most critical medium for scholarly publication in many disciplines—especially in the humanities and certain social science areas. Many such works are published by university presses—scholarly publishing enterprises run by universities that publish scholarly works whose audience will be primarily scholars.

Much significant work is done by doctoral students, whose research is documented in their dissertations. Much of this work is later disseminated in other forms—presentations at meetings, articles, monographs—but some is never published and archived in these kinds of venues, and in such cases the dissertations are the major archiving resource for the research results.

Most of the focus of this paper is on traditional late-stage scholarly communication—i.e., on publication. The main point is that we are really looking beyond books and journals as we know them to media of the future that are not known. The concept of the "book" or "journal" may be last century—largely a function of changing media, but also a function of changing markets, changing library practice, and other issues. A recent report from the Association of University Presses (Withey, et al., 2011) says: "e-books lend themselves to new forms of distribution…breaking down the concept of the "book" in favor of packaging content…" Similar thoughts apply to journals, which as noted above have moved far toward digital format and distribution technologies. Even more transformative, however, is the question of why a digital "journal"

could not publish modular "books" in the form of "article-like chapters". Such modules could include much more than text—e.g., audio, interactive elements, graphic content, and manipulative elements (e.g., simulations). Are books and journals as we know them the buggy whips of the twenty-first century?

**Principles of Scholarly Communication**

One of the most important issues for us to address as we navigate the future of scholarly communication is the deep and compelling culture of higher education that frames the principles of scholarly communication. These are issues for which faculty and others in the scholarly world have deep ethical values. That is, we must be clear about the desired benefits and effects of scholarly publishing in this broad scholarly context.

Perhaps the most compelling principle behind faculty members' ideas about scholarly publishing is that there should be open communication with regard to content (but also see Harley, et al., 2010, p. 13, on disciplinary differences in openness). Scholarly results MUST be freely accessible to the scholarly world—and to those who may use such results for applications in business, government, or other domains. We will come back to this matter in several ways. One that is of increasing concern at the present time is that there are now many limitations on open communication that arise from security and commercialization interests. There seem to be three main areas of concern.

Issues of protecting intellectual property (IP) have become increasingly important and restrictive in the past few years, especially as universities' commitment to economic development has increased, as has commercialization of IP created by the universities' researchers. The issue is closely related to the issues underlying the Bayh-Dole act of 1980, which recognized that if IP cannot be protected, much will never come to public good, since the development costs cannot be justified without such protection. Thus, as IP development and commercialization has become a central part of universities' mission, the implications for free and open communication about research results have been negative. There is no clear and simple answer to this conflict, but it is one that must be recognized.

As national security issues became ever more compelling—especially following 9/11—the transfer (intentional or otherwise) of certain IP with security implications became very sensitive. Export control policies that limit the risk of security-sensitive IP leaks have now become a significant limit on open communication of research results. But export control now goes far beyond the open communication of research results—for example, to limiting the presence of international students in a lab with instrumentation with potential security relevance. Moreover, students who work in labs and/or with professors whose work has export control sensitivity may find that their theses or dissertations are embargoed. Since export control began, the grounds for limiting communication about and/or participation in research has expanded to include the transfer of information about IP that has significant economic implications: e.g., transferring information about an economically transformational technolo-

gy to another country through scholarly publishing, international collaboration of scholars, or having international GAs in a lab.

Classified research has long been a hard limit on scholarly communication and continues to be a significant issue.

Another critical element of scholarly communication is the validation of the quality of the work (Harley, et al., 2010, p. 10, 12). There are actually many concerns about the increasing number of low-quality publications, some from self-publishing, vanity "academic presses", and reduced quality of peer review. In any case, the main quality assurance mechanism for scholarly publishing is peer review, the vetting process by which articles, books, presentations at meetings of scholarly organizations, proceedings, and other modes of scholarly communication is done. Basically, the idea is that highly qualified peers of the researcher whose work is being validated read the work and express their judgment on the validity and impact of the work. Some of this kind of peer review occurs very early in the scholarly process—e.g., for grant proposals—but much comes very late in the process after the research is completed and being considered for publication—one might say at the "archival stage." This is a very important issue—that work archived in prominent (i.e., respected) "publications" has been vetted and can be considered "reliable."

There are some significant issues about such peer review. One is the question of whether it is entirely consistent with the principle of open scholarly communication. We know that the most respected publishing venues tend to be conservative, raising the question of whether really innovative and high-impact research will be "accepted" for publication. That is to say, the mindset of peer review is restrictive. This stems from the idea that for journals and presses, the lower the acceptance rate, the higher the perceived quality is seen to be. Peer review thus becomes a central element of assessment for academic ratings of individuals' work, and it is thus critical for promotion and tenure, hiring, awards, NAS membership, and other quality assessment circumstances. It also becomes a core element for ratings of academic programs and of universities. Like so many other elements of scholarly publication, peer review becomes a tradeoff: open access, real innovation, and validation of quality. And, again, there are no clear and simple answers (Harley, et al, 2010, p. 22; Harley, et al., 2011).

In any case, peer review in its current form is a costly element of publishing, both for journals and books. It is not that the peer reviewers get paid to do their reviews; rather, what is costly is the vetting of the potential reviewers and handling the clerical and evaluative function after the reviews are received. And, more important, it is not just the works that are published that are peer reviewed, but for the most highly regarded venues, the number of reviews greatly exceeds the number of publications—the restrictive criterion for stature being a driver of such dynamics. There have been some discussions of alternative methods of peer review—e.g., post-publication peer review—but there is at present no consensus on good practice for the future.

There has also been a great deal of discussion of various models of open access and the role of peer review (Withey, et al., 2011, p. 14-15). One model of open access keeps the traditional model of peer review in place. "Open access" is not the "acquisition" or "acceptance" model, but the model for access to the ultimately peer-reviewed work that appears. What is "open" is access to the "publication." But some models extend the "open" matter to acceptance, thus addressing the question of whether peer review, which is inherently conservative, violates "openness" of communication; but this practice is not so clearly in line with the idea that the "published" research should be validated before being open to all potential readers.

Another issue that significantly impacts issues of prestige, rankings, and access by scholars to scholarly publications is interdisciplinary scholarship (Harley, et al., 2010, pp. 7-8, 15-17). Much of the highest impact research today is interdisciplinary. But, that said, most of the high-prestige publishing is extremely discipline centric, and thus conservative, in the sense that the works that are "accepted" are at the center of the disciplines. Although the interdisciplinary work tends to be more high impact, the prestige of journals or presses accepting such work tends to be lower than those that are highly restrictive and disciplinary centric. Therefore, the journals with a high "impact factor" tend to be those that are disciplinary centric. This is a daunting issue for scholars doing high-impact interdisciplinary work, since it very negatively impacts their prospects for promotion and tenure, for hiring, and for other processes that involve evaluation of scholarly work. Accordingly, it poses significant disincentives for taking on risky, interdisciplinary, and potentially high-impact research. And accordingly, interdisciplinary work tends to impact rankings of departments, programs, and institutions in complicated ways (e.g., if many faculty are doing high-impact interdisciplinary work not published in high impact journals).

In many ways, the most important function of scholarly publication is the archival function. As we have indicated, most of the scholarly communication occurs before publication of the results. Leading researchers in a given area are generally in close, constant contact with peers. They exchange information at conferences, symposia, and in contexts as informal as having a beer together. But journals and monographs archive the research results for the long future. The use by future scholars may be WAY in the future. In fact, few pieces get much attention in the future, but some do…and may be critical for very high impact scholarly success in the distant future. But this kind of archiving function raises some significant questions about the long-term viability of the digital publications as an archival function. For example, there does not seem to be a viable plan to migrate the staggering amounts of "archived" digital data if significant new technologies emerge. Can we really migrate a trillion petabytes of data when the current technologies are replaced?

Finally, with respect to the principles of scholarly communication, it is critical that the educational function be

considered. Articles and monographs are critical resources, especially for advanced (e.g., graduate, professional) education. Advanced undergraduate and graduate students' resources are to a large degree reading the results of earlier scholarly research. But there are many practical and legal challenges to the use of such materials in digital format. For instance, imagine a multi-campus graduate program in which courses from several campuses serve students of other campuses. If the licensing of journals restricts access to members of the home-campus community, students at the other campuses will be unable to access significant educational materials. We at MU have experienced this problem. So, there is a tradeoff for whether the library pays significantly higher licensing fees for a journal to provide broader access or whether it subscribes to more journals for the campus community.

Also in the educational area, there are many questions about how new modes of "publication" fit in the picture with regard to IP and other issues. It is a certainty, at least in my view, that we will see dramatic changes in the formats of educational materials. These issues will create daunting questions with respect to IP rights of faculty creating the materials. But even more importantly, it will create major problems about cost of very high-impact materials created by innovative publishers (e.g., simulations by Disney or Microsoft developed at the cost of tens of millions of dollars, useful for hands-on learning in small learning groups). We have to ask if textbooks as we know them will exist in the future. Why would we not move to flexible modular digital "chapters" that can be aggregated like "articles" or "book chapters" and linked to very high-level technologies that facilitate hands on learning—e.g., a simulation of forest ecosystems, or of businesses, or of urban development?

**Practical questions**

There are many practical/operational questions about the future of scholarly communication—especially scholarly publishing. There is a lot of discussion of the fiscal "solution" being digital publishing. The fact is that the cost of digital publication is only about 25% less than the cost of paper publication, assuming that the same level of peer review and other functions are similar. The bottom line: the cost of digital publication is nearly as expensive as paper publication. Savings is not the issue for digital publication. The model for revenue generation for digital journals is clearer than for other publishing media such as monographs, but it's not clear that it is sustainable in any case. An important question is whether digital "journals" may be repositioned in scholarly publishing in areas where monographs have been dominant? Could a "journal" do "monographs"? What would be the effects on monograph-based disciplines…and on university presses?

There are other challenges to the fiscal viability of university presses. For instance, the sales volume for the typical monograph over about 10 years fell from about 1,500 to 600-700 copies (e.g., Lamm, 1996, p. 136)). Another limitation on revenues for University Presses and a challenge to the archival function is the "library purchase on demand" model, in which libraries purchase books only af-

ter there is a demand (i.e., a client request for access) for the book. This kind of purchasing is becoming very common in library practice, and it has potential to significantly impact the business model of university presses.

Whatever the future path, there will be unexpected consequences. There are very complicated implications for the academic principles of scholarly communication—e.g., peer review is called into question, and the potential implications for promotion and tenure and hiring processes can be severely impacted (Harley, et al., 2011). The issues of "impact factors" and related matters were discussed above. But even more important, many other matters affect the appearance of impact, prestige, and stature of publications and, therefore, feedback on program stature and institutional rankings. A more concrete issue: there is a perspective by which digital subscriptions in libraries limit access to scholarly results. One doesn't need to "sign in" to take a paper volume off the shelf in the Library; but if one wants access to the digital journals, one has to be a "member" of the "organization" (e.g., university) to be able to "sign in" to get access to the on-line material—a condition of the licensing. Of course, licensing could be done in a much more expensive and inclusive mode…but what is the tradeoff with the number of subscriptions the library can sustain?

**University of Missouri Press**

A compelling example of the urgency and the complexity of the scholarly publishing issues has been captured by the national—even international—push-back that occurred when the University of Missouri System President announced that the University of Missouri Press would be closed to save the $400,000 per year subsidy (and additional deficits after the subsidy) that were being paid. It was not just the University of Missouri faculty's outrage—even more, it was a national response (outrage is perhaps too strong, but maybe not) that the University would abandon its obligation as a major research university to support one of its most important functions: scholarship (see a small sample of the press coverage: Singer 2012, Williams 2012, and Eligon 2012). The President's decision was driven by the daunting fiscal challenges facing the University of Missouri as state appropriations dropped and costs continued to rise. What he perhaps did not realize was the degree to which scholarly communication is central to the research mission of higher education. It is not surprising that research journals and university presses are not self-sustaining, but their function, as outlined above, is critical to the scholarly mission.

Fortunately, there had been very productive discussions for several months prior to the President's announcement, involving both System and MU Campus people—faculty, administrators, staff, and others—about the daunting challenges facing university presses. And when President Wolfe's announcement occurred, there was a rich set of discussions on which the MU campus could build a new vision for the University of Missouri Press moving to the Campus. Basically, the ideas were based on two main premises. First, the Press must continue its main function of disseminating and archiving major

scholarly outcomes. From the point of view of authors, it would not look different than the previous press model in the short term (e.g., would continue print publications, strong peer review in the traditional sense, and would do marketing at least as effectively as in the past). More importantly, it would be embedded in the campus academic environment in a way that was not possible when it was a "System" function. Namely, it would have both an instructional and research connection that would position it well in the rapidly changing environment for scholarly communication.

The details of this academic engagement have not yet been worked out, but the broad vision is clear. On the one hand, our international academic program strengths in Journalism, Library Science, Creative Writing, and other areas would develop an academic program (perhaps a certificate program) that would prepare students for careers in the volatile world of scholarly communication. One element of such a program would be internships and student employment at the Press, with students mentored by editorial and other staff who have a faculty function. On the other hand, we could build on our strengths in new media (our world-class Journalism program and the Reynolds Journalism Institute), Library Science, Creative Writing, and the University Libraries to do cutting-edge research on how scholarly communication is changing. From this latter point of view, MU has potential to become a world leader in not just understanding where scholarly communication is going, but in shaping a major university press that could become a model for others as these complex dynamics play out.

From this perspective, the Press becomes an important entity much like an internationally prominent research lab or center. It is embedded in the broad academic mission of the University, providing a venue for strong graduate student learning experiences at the same time as it is linked to a major research program. There are many models for such engagement at the University: science research labs, clinical programs, the Law Review, Medical Education, and others. We believe this is a strong, forward-looking model for scholarly publishing that may become a model for the future.

**Bottom Line**

The issues involving scholarly communication are very complex. There is no clear, simple answer. A key issue is that we must identify the unintended consequences of change—and the only certainty is that dramatic change will occur. We know that media technologies are impacting scholarly publication in profound ways. Costs of digital communication will be high—not much less than paper. A sustainable revenue stream must be found. In any case, we must mitigate the unintended consequences such as limiting access as a condition of library subscriptions. Open access, peer review, and "impact factors" are hard to reconcile. Our values about "free and open communication" and "peer review" and "high impact/prestigious" research are contradictory with one another. Monographs, journals, and other media forms will change in the digital environment. We are likely to see entirely new forms that

do not align with journals and monographs—the "buggy whips" of scholarly communication.

There are no clear answers. We need to have deep and open dialogue. We need some models for new models for scholarly communication—perhaps the new model for the University of Missouri Press. And we need to understand that the only thing we really know is that the current system is not sustainable.

**Acknowledgements**

I would like to thank participants in the Merrill Retreat for their useful feedback on a very early, oral version of this paper. I would also like to thank Don Lamm (former CEO of Norton Publishing and former member of several University Press boards) and Lynne Withey (former director of the University of California Press) for their extraordinarily helpful comments on a near-final draft of this paper. They bring very different and very complementary perspectives to the issues of scholarly communication and changing media technologies.

**References**

1. Eligon, John. "Plan to Close University of Missouri Press Stirs Anger." New York Times, July 18, 2012.
2. Harley, Diane, and Sophia Krzys Acord. Peer Review in Academic Promotion and Publishing: Its Meaning, Locus, and Future. 2011. Berkeley, CA: Center for Studies in Higher Education.
3. Harley, Diane, Sophia Krzys Acord, and Sarah Earl-Novell. Assessing the Future Landscape of Scholarly Communication. 2010. Berkeley, CA: Center for Studies in Higher Education.
4. Lamm, Donald. "Libraries and Publishers: A Partnership at Risk." Daedalus, Fall, 1996.
5. Singer, Dale. "New Version of UM Press will Involve Students, Research." St. Louis Beacon, July 16, 2012.
6. Williams, Mara Rose. "After Outcry, University of Missouri Presents New Plan for Press." Kansas City Star, July 17, 2012.
7. Withey, Lynne, Steve Cohn, Ellen Faran, Michael Jensen, Garrett Kiely, Will Underwood, and Bruce Wilcox. Sustaining Scholarly Publishing: New Business Models for University Presses. 2011. New York, NY., The Association of American University Presses.

# Smart Infoware: Providing University Research Stakeholders Soft Power to Connect the Dots in Information Haystacks

**Chi-Ren Shyu,** Director, Informatics Institute; Shumaker Endowed Professor of Informatics, University of Missouri

While resources have been allocated to build computing infrastructure everywhere in the nation, the value of infoware to assist the university scientific communities has been underestimated, or even ignored. Those unsung heroes, who develop infoware to dig into the information haystacks and provide to the research community needle-sized up-to-date knowledge, deserve recognition, since without their innovative work scientific discovery would not be able to move quickly from bench to bedside in healthcare or move from greenhouse to dishes in agronomy and food production. Moreover, the organization and coordination of available infoware are needed to leverage regional talents to equip researchers with soft power as opposed to the hardware-based computing muscles.

**Informaticians and Infoware**

Informaticians often identify research problems and then model potential solutions in a mathematical or computational way in order to streamline the knowledge discovery process so that costly experiments will be minimized. I would like to use University of Missouri Informatics Institute (MUII) to demonstrate the development process, application, and potential of infoware developed by the informaticians. The history of MU informatics research began in the 1960s when Dr. Donald A. B. Lindberg, the Director of National Library of Medicine, pioneered the application of computer technology to healthcare. In 2012, there are 42 core faculty members from 14 departments and 7 colleges/schools who contribute to the curriculum development and research/outreach activities to support a doctoral program hosting 35 PhD

Figure 1. In the personalized medicine era, physicians will need to hash through complex information for accurate and personalized care.

students. The informatics community at MU continuously develops infoware to serve the worldwide research community in three major areas – bioinformatics, health informatics, and geoinformatics. This infoware has played significant roles in handling the ever- growing size of data in genomics, proteomics, biometrics, and imaging technologies. As shown in Figure 1, future healthcare will need to connect the dots from sensor data for homecare patients; understand various image modalities; recognize protein structure data for drug reactions; and comprehend personalized sequence data. It becomes more and more challenging when the goal of "$1000 genome" will soon be reached. It is a known joke in the industry that "sequence a whole genome overnight and take months to analyze it." The bottleneck lies in the computational inability to process the data in each lab without shipping the data through ultra-high-speed network to a high performance computing facility

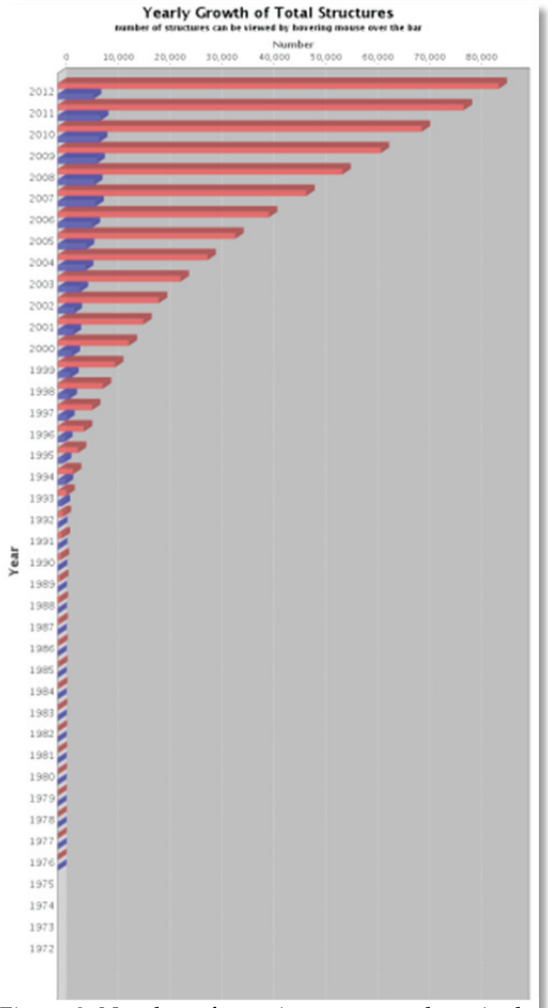


Figure 2. Number of protein structures deposited in the Protein Data Bank (Picture from the PDB database).

**Bioinformatics**

**Figure 2** shows the growth trend of protein structures in the Protein Data Bank (PDB). Without smart infoware to analyze the data, the financial burden to purchase larger computer clusters becomes bigger and unmatchable to the growth of information. Examples of infoware developed by MU bioinformaticans include:

SoyKB (http://SoyKB.org): Soybean knowledge base is a comprehensive all-inclusive web resource for soybean research. It provides an infrastructure to handle the storage, integration, and analysis of the gene, genomics, EST, microarray, transcriptomics, proteomics, metabolomics, pathway and phenotype data.

ProteinDBS (http://ProteinDBS.rnet.missouri.edu): A Web server designed for efficient and accurate comparisons and searches of structurally similar proteins from a large-scale database. It provides worldwide users a speedy search engine for global-to-global and local-to-local searches of protein structures or substructures.

DOMMINO (http://dommino.org): A comprehensive database of macromo-

lecular interactions that includes the interactions between protein domains, interdomain linkers, N- and C-terminal regions and protein peptides. This infoware provides a flexible search for interactions and visualization tools to navigate the search results.

PhenoDBS (http://PhenomicsWorld.org): A web-based infoware that provides the genomics community with a computational framework to study complex phenotypes using techniques such as visual content management, semantic modeling, knowledge sharing, and ontology customization.

All the infoware shares the same goal, which is to provide open-source tool access to the entire scientific community with speedy search from large-scale and complex data sets that normally cannot be organized and processed without high performance computing.

**Health and Medical Informatics**

Billions of dollars have been invested to drive innovations in health care technologies, such as the Nationwide Health Information Network (NHIN), to enable health information exchange with secured, simple, and seamless infrastructure. This provides physicians an environment with meaningful use of electronic health records, and Centers for Medicare and Medicaid Services (CMS) has the Health Care Innovation Awards to implement the most compelling new ideas to improve the quality of the care as well as reduce the cost for citizens enrolled in Medicare, Medicaid, and Children's Health Insurance Program.

One example of infoware developed by MU health/medical informaticians is the informatics framework for the American Lymphedema Framework Project (http://ALFP.org). This project provides an operational cyber infrastructure that collects, organizes, and disseminates up-to date lymphedema (LE) information. It includes the development of a cyber framework for inter-institutional lymphedema research activities; the integration of informatics tools to provide an on-line summary of concurrent LE studies; and the creation

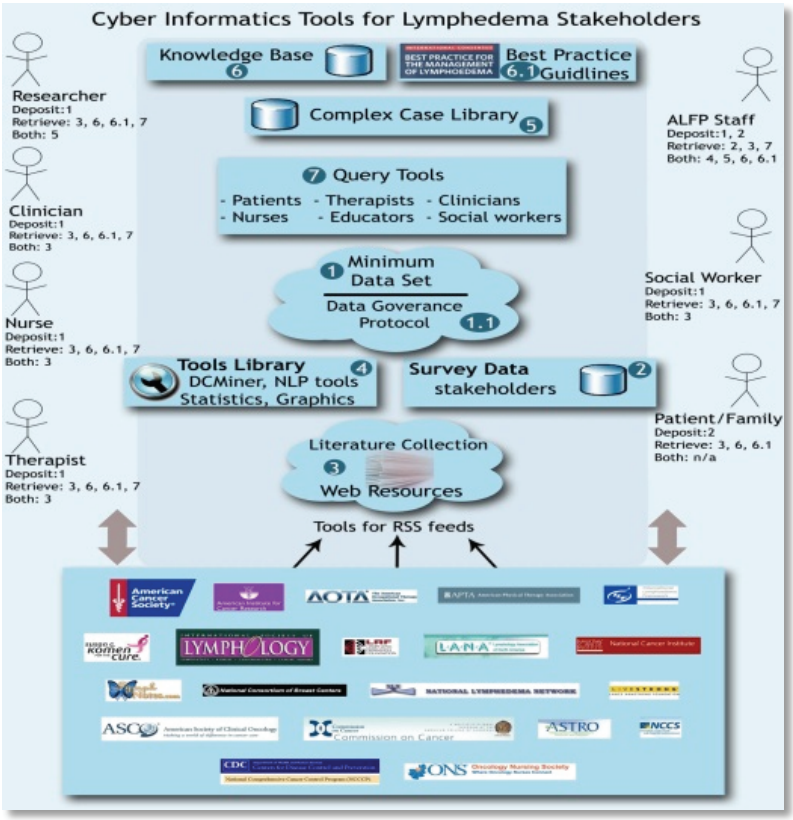



Figure 3. A medical informatics infrastructure to provide infoware to serve an international community in lymphedema.

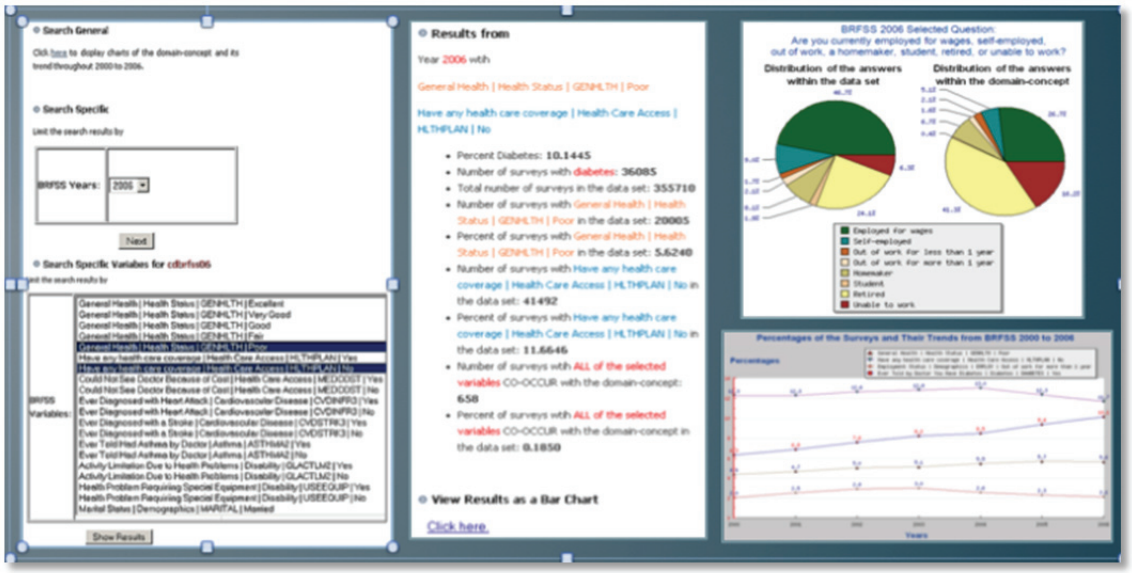


Figure 4. Infoware for clinicians to map a single patient's profile to high-risk populations for real-time alerts and disease management.

of a web portal for LE stakeholders with user-specific query methods. Infoware developed for this project have been targeted to serve users from all ranges of backgrounds, such as patients, therapists, physicians, researchers, etc. In addition, mobile apps, such as the iPhone App "Look4LE" as shown in Figure 3, are also developed to provide information access to patients.

Another infoware example in health and medical informatics is mapping tools for researchers and health professionals to connect a specific patient with public knowledge through public health information systems, such as cancer registry and Behavioral Risk Factor Surveillance System (BRFSS) data from the Centers for Disease Control and Prevention (CDC). Applying mining tools to extract clinically significant patterns for various demographic and geographic populations, researchers and clinicians can compare a patient's profile with the trends of healthcare related activities for high-risk patients from the public information systems. This infoware can provide real-time alerts for patient's disease self-management and for clinicians to provide proactive cares.

**Geoinformatics**

More and more data are expected to be geo-coded when the number of data acquisition devices with embedded GPS functions increases. Geographic Information System (GIS) analytics can be used to provide the research community the ability to analyze geospatial information for decision-making. It has been postulated that GIS could provide distribution of available information resources in order to bridge the gap between at-risk patients and access to therapists and treatment centers for chronic diseases. GIS-based infoware is utilized to analyze patients and chronic disease resources as a geographical representation in order to identify associations between them. Additionally, we sought to provide information to inform policy-makers and educators in order to assist with resource development, prioritiza-

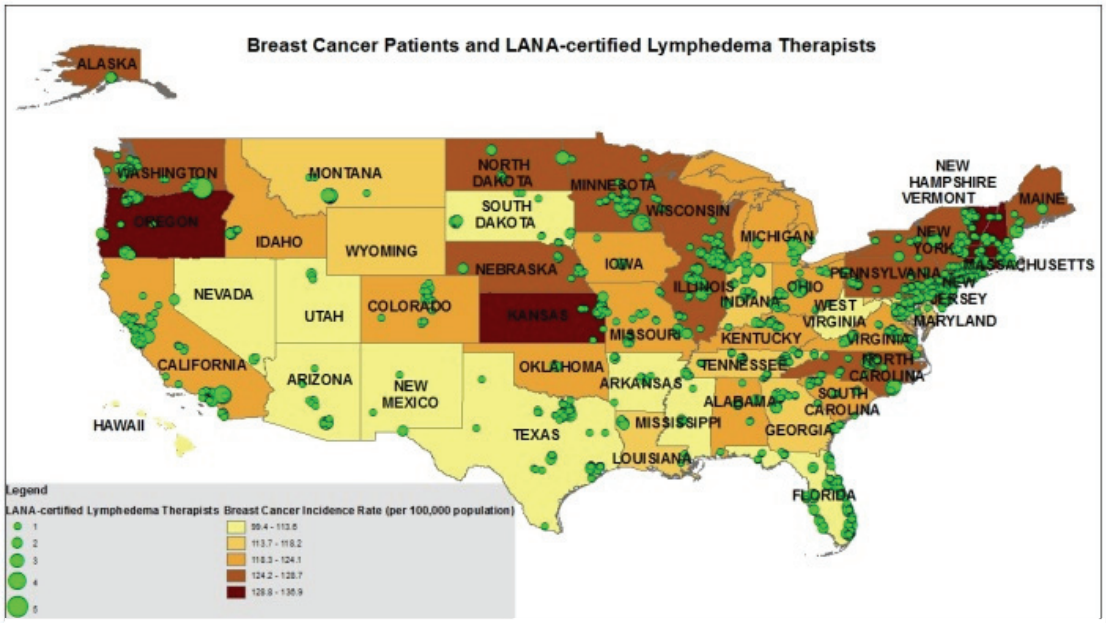



Figure **5**. A GIS-based view of available resources for breast cancer survivors who might suffer from certain chronic diseases.

tion, and allocation. Figure 5 shows available certified therapists for lymphedema patients.

**Scenarios Involving Multiple Informatics Areas**

One example of infoware that involves techniques from both bioinformatics and geoinformatics is the development of tools that can allow plant genetic researchers and breeders to search crop plant disease phenotypes in real-time using genetic, phenomic, and geospatial information. The technology has been tested with the maize community to study mutants and diseases

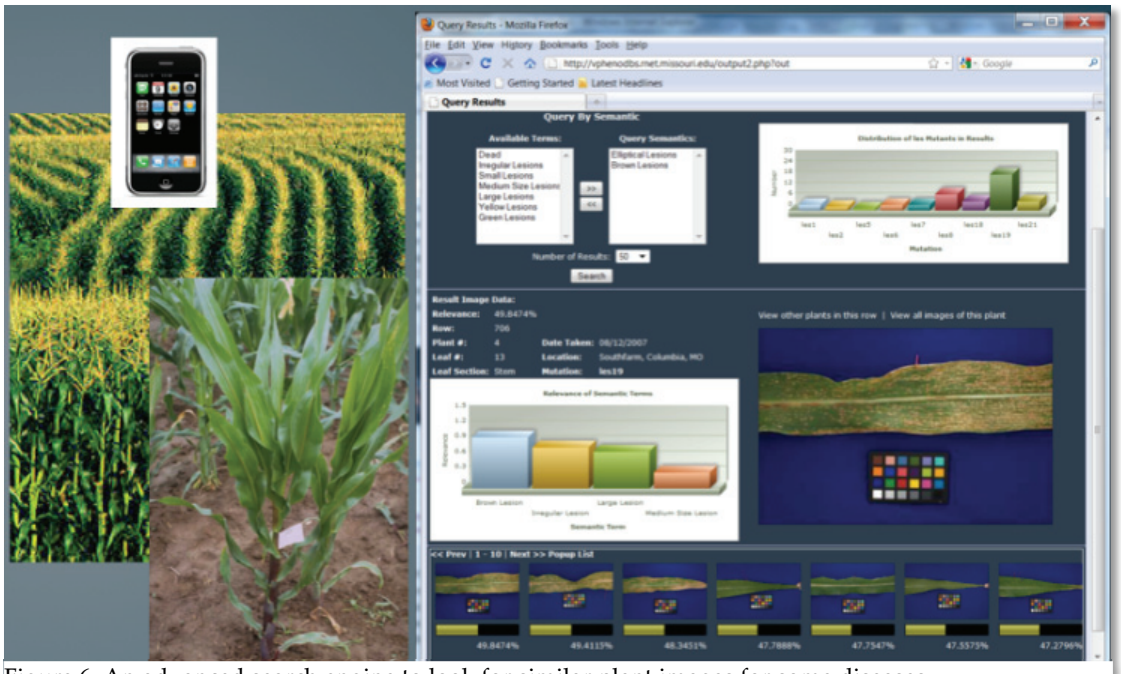



Figure 6**.** An advanced search engine to look for similar plant images for some diseases.

to streamline the process of plant morphology, ecology, and phytochemistry research. Such infoware is unique in the sense that it provides GIS-enabled query by phenotype images, query by semantics in mutant and disease descriptions, and other customized complex query methods to assist plant researchers studying the underlying effects of genetics and environmental factors on physiology. Figure 6 shows the advanced search tools that will enable future research in plant genetics by fostering cross-institutional study of complex traits using unique infoware developed by informaticians.

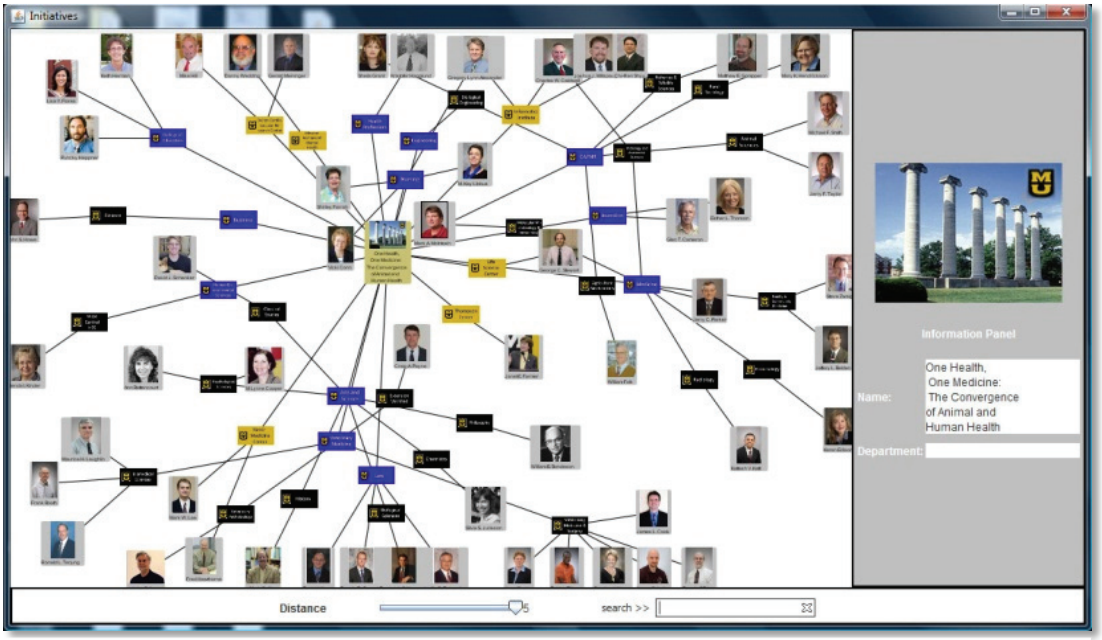


Figure 7. Research network for "One Health, One Medicine" initiative at MU.

**Conclusions**

Infoware has been developed independently by colleagues in Iowa, Kansas, Missouri, and Nebraska. However, most informaticians are unaware of these developments in their surrounding institutions. Thus it is unlikely to expect researchers in other fields to understand the regional talents that can greatly enhance their research using the existing infoware. Therefore, it is necessary for infoware developers from the region to meet and put together an infowarehouse for tool sharing and education.

Moreover, a research social network which is searchable by university researchers and industry partners is also needed for the region. This linkage of researchers may consist of co-authored publications, collaborative proposals for extramural grants, student committee memberships, national/international committee services, etc. Challenges occur mainly on data collection and quality control when multiple institutions are involved for this talent knowledge base. Figure 7 illustrates an example of a research network for researchers who have collaboration records in the area of the One Health and One Medicine Initiative under the Mizzou Advantage program. To make the network valuable, an ideal infoware for such type of network should provide talent search as powerful as leading search engines, such as Google™, and recommend researchers as thoughtful as popular shopping sites, such as Amazon.com™.

# Finding Subatomic Particles and Nanoscopic Gold with Open, Shared Computing

**David Swanson,** Director, Holland Computing Center, University of Nebraska

Modern data intensive research is advancing knowledge in fields ranging from particle physics to the humanities. The data sets and computational demands of these pursuits put ever-increasing strain on University resources. The history of resources at the Holland Computing Center (HCC) at the University of Nebraska (NU) reflects these dramatic changes. The need to constantly evolve and grow resources economically requires careful strategy and provides incentive to combine and share resources wherever possible. The remainder of this paper provides a brief overview of HCC and outlines some recent highlights and strategies that have emerged over time.

## HCC Overview

### Facilities

HCC has two primary locations directly interconnected by a pair of 10 Gbps fiber optic links. The 1800 sq. ft. HCC machine room at the Peter Kiewit Institute (PKI) in Omaha can provide up to 500 kVA in UPS and genset protected power, and 160 ton cooling. A 2200 sq. ft. second machine room in the Schorr Center at the University of Nebraska-Lincoln (UNL) can currently provide up to 60 ton cooling with up to 400 kVA of power. A cooling upgrade to match the maximum power load is in process. Both locations have 10 Gbps connections to Internet2 in addition to a pair of 10 Gpbs links connecting the machine rooms.

HCC's resources at UNL include two distinct offerings: Sandhills and Red. Sandhills is a linux "condominium" cluster provided for general campus usage with 1500 compute cores interconnected by low-latency Infiniband networking. A recent extension provides a higher RAM/core ratio. The largest machine on the Lincoln campus is Red, with over 3100 cores interconnected by less expensive, but also higher-latency, gigabit Ethernet. More importantly, Red serves up over 1.2 PB of storage. Red is integrated with the Open Science Grid (OSG), and serves as a major site for storage and analysis in the international high energy physics project known as CMS (Compact Muon Solenoid).

In late May of 2009 the University was donated Firefly, a 1152 node, primarily dual-core Opteron cluster which became the largest resource in the University system. It is connected by Cisco SDR Infiniband and supports 150 TB of Panasas storage. Originally capable of 21.5 TFlops, it is located at PKI, and is nearing retirement. The latest cluster named Tusker is also located at PKI. Tusker offers 6,784 cores interconnected with Mellanox QDR Infiniband along with 360TB of Terascala Lustre storage.

Each individual node contains 64 compute cores with a total of 256 GB RAM.

**Campus Grid**

HCC maintains a Campus Grid that is able to transparently submit large batches of jobs first to the campus, then to Big 10 peers Purdue and Wisconsin, and finally to the OSG. Over 4M cpu hours have been processed via the campus grid framework over the last year. Processing is occurring significantly in Nebraska, but also at Fermilab, Wisconsin, Purdue, Michigan and many other locations as far away as Caltech and Brazil. This method of computing is also able to access resources on the Amazon EC2 Cloud. These have been minimally utilized due to the fact of current pricing but the spot pricing option may become sufficiently affordable soon. Recent studies by the DOE concerning storage have continued to conclude that Cloud resources, even if capable of the scales demanded by data centric science -- not yet confirmed -- are currently far too expensive compared to locating resources at Universities and Labs.

**Open Science Grid**

*"The Open Science Grid (OSG) advances science through open distributed computing. The OSG is a multi-disciplinary partnership to federate local, regional, community and national cyberinfrastructures to meet the needs of research and academic communities at all scales."*
[https://www.opensciencegrid.org/bin/view/Documentation/WhatIsOSG]

HCC's relationship with OSG has grown steadily over the last several years, and HCC is committed to a vision that includes grid computing and shared national CI. Recently, an HCC VO was formed, and researchers affiliated with HCC are now able to run on all HCC resources as well as resources across the globe through OSG protocols. The experience with OSG to date has proven to be a great benefit to HCC and NU researchers, as evidenced by the HCC VO's usage of over 11 million cpu hours in the last calendar year.

**Data Intensive High Throughput Computing @ HCC**

**US CMS Tier2 Computing**

HCC is a substantial contributor to the LHC (Large Hadron Collider) experiment located at CERN, arguably the largest cyberinfrastructure effort in the world. Researchers affiliated with the LHC are probing previously unreachable energy regimes to expand our understanding of the very fabric of the universe. In particular, they have been searching for the Higgs, or "god," particle, which will help to explain why matter has mass.

One of the experiments at the LHC is CMS. 30 PB of disk space (in an archival, "write once, read many" mode) and 38 PB of tape were required for the complete analysis of just the first year of LHC data. The storage and processing requirements are at unprecedented scales for scientific computing. In the US, the underlying grid middleware is provided by the OSG.

HCC operates one of the seven CMS "Tier-2" computing centers in the United States. Tier-2 centers are the primary sites for user analysis of CMS data; they host datasets that are of interest to physicists, who submit jobs over the grid to process those datasets. They are also the primary sites for generating collision simulations that are used by physicists throughout the experiment; these jobs

| | 1999 | 2002 | 2005 | 2009 | 2012 | Increase |
|---|---|---|---|---|---|---|
| Personnel | 2 | 3 | 5 | 9 | 13 | 7x |
| WAN band-width (gbps) | 0.155 | 0.155 | 0.622 | 10 | 30 | 200x |
| CPU cores | 16 | 256 | 656 | 6956 | 14492 | 900x |
| Storage (TB) | 0.108 | 1.2 | 31.2 | 1200 | 3250 | 30,000x |

are submitted over the grid by a central team of operators. The Nebraska Tier-2 currently has 3000 processing cores and hosts 1200 TB of CMS data with redundant replication.

"I think we have it. We have discovered a particle that is consistent with a Higgs boson." This quote from CERN Director-General Rolf Heuer was the remarkably understated announcement that a decade-long effort by an international team of thousands of scientists using arguably the largest cyberinfrastructure environment in world history had made a major breakthrough. Although a myriad of fine details remain to be determined, something that is "consistent" with a Higgs field -- and a mechanism that gives rise to mass for all matter -- has been confirmed.

Movement of data sets of these sizes requires the use of high performance networks. UNL has partnered with other regional Universities in the Great Plains Network (GPN) to share the cost of obtaining and maintaining connectivity to Internet2. The Tier2 site at HCC routinely moves data at rates approaching 10

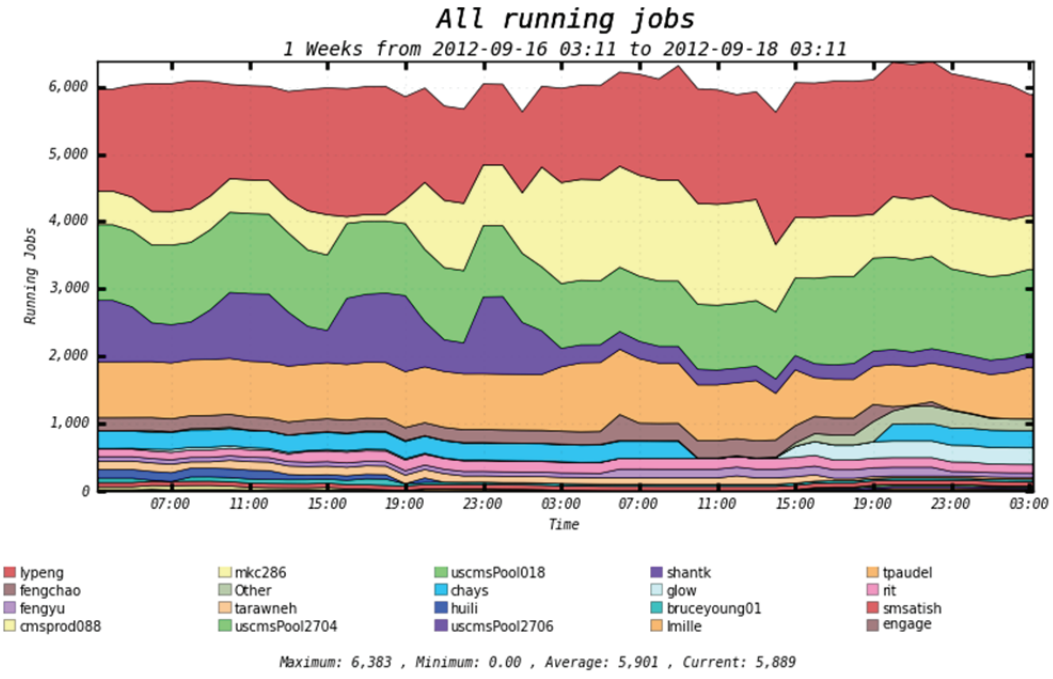



**Figure: Mixture of Jobs at HCC**
The figure above shows a 2-day window of running jobs at HCC. The salient feature is that the 6400-cores available during this window are efficiently utilized. The green swatch near the middle is composed of grid jobs, surrounded by layers composed of more traditional HPC jobs.

GigaBits per second to and from other LHC collaborators, often via this shared Internet2 link.

**HPC @ HCC**

While shared high throughput computing (HTC) is a major thrust at HCC, it should be emphasized the majority of computing done at HCC is still traditional high performance computing (HPC), since this is what most HCC researchers currently need to advance their research. The fact that HPC coexists so well with HTC at HCC is strong evidence this model of shared computing can be extended to other locations. Here we briefly highlight several recent high profile findings from the Zeng group at UNL that relies upon HPC.

This summer Xiao Zeng published three high-profile papers, all of which required complex computations completed with substantial usage of HCC. All told, Zeng's group used over 11 million cpu hours over the last year. Zeng and colleagues created a model potassium channel -- a nanoscale sieve -- much like found in almost all living cells. Their computations helped to guide the experimental synthesis of this assembly, which was published in Nature Communications in July.

The next month Zeng's group published in the Journal of Chemical Physics a novel 7-coordinate form of carbon CTi7(2+). Finally, Xiao Zeng published in Science an "ordered amorphous carbon cluster" so hard it can dent diamonds, the hardest known substance. These were examples of computational muscle merged with chemical imagination. HCC was built to accelerate this type of success; HTC has not hindered these advances.

Overall, HCC facilities usage is composed of both Nebraska research and international collaboration. A total of 10 Million cpu hours are used monthly at HCC between our various machines. Top usage includes groups belonging to Xiao Zeng (UNL Chemistry), Anthony Starace and Stephen Ducharme (UNL Physcis), Clint Rowe (UNL Earth and Atmospheric Science), Renat Sabirianov and Wei-Ning Mei (UNO Physics). These represent traditional computationally intensive processing. Interleaved with these users are various Virtual Organizations (VOs) from the Open Science Grid (OSG), led by CMS, which is a direct collaborator with HCC. Others include local "grid" and "HCC" usage, which are direct usage from UNL scientists, including bioinformatics and mathematics. OSG usage is opportunistic, and only uses HCC machines when capacity is available due to fluctuating demand from NU researchers. While it is expensive to operate computational hardware, the annual operational expenses of HCC are dwarfed by the amount of depreciation of the hardware itself.

**HCC changes over the last decade**

HCC resources have changed dramatically over the last 13 years. Data prior to 2009 is from the previous organization known as the Research Computing Facility (RCF). The following table gives estimates of the amounts of various cyberinfrastructure resources from several years during this time period.

While details will be different for other locations, the qualitative trends are commonly observed. The relative increase in storage capacity at HCC far surpasses that of the other categories. Data sizes are increasing 33 times as fast as the number of CPU cores, and 150 times faster than the WAN bandwidth needed to move the data. For comparison, it takes a month to move 3 PB of data at a sustained 10 gbps. It is thus

not surprising to see 100 gbps initiatives in the research community. Power usage has increased more slowly than the other hardware increases, although precise numbers are difficult to obtain prior to 2009; it has been roughly constant over that time.

**Observations and Guidelines**

The experiences listed above lead to several guidelines that aid in the management of HCC.

*We need to use what we buy*. These pieces of cyberinfrastructure are linked, but improve asynchronously. It is exceedingly difficult, if not impossible, to future-proof any hardware investment. Depreciation, on the other hand, is immediate. Leasing, such as purchasing compute time and storage from Amazon Web Services, is not yet cost effective in most cases. Especially because hardware improves dramatically over time, one can increase the return on investment by buying at regular intervals. This is manageable if one purchases on an annual basis. Space, power and cooling infrastructure tends to outlive any particular hardware purchase regardless.

*We should share what we aren't using.* By giving opportunistic access to researchers from other organizations, we do not relinquish control of our local resources. We do, however, serve a greater good within the broader scientific community, and utilize otherwise squandered idle resources which are rapidly depreciating. In this process, we have the potential to contribute and gain collaborators. These collaborators may in turn share their resources with us, allowing us to burst usage to greater scales when a situation demands it. There are always many more resources off campus than those currently possessed on campus.

*A data deluge is upon us.* While other surrounding hardware is increasing at remarkable rates, the amount of data storage capacity demanded is currently outstripping processing and data transfer advances.

*Expert support is essential.* If we build it, they may not come. Support personnel can be increased at rates much lower than that of the hardware they maintain and facilitate. By acquiring hardware incrementally and regularly, one buys time for the user community to be trained to effectively exploit resources.

*Links to regional and national infrastructure are critical.* The scientific advances that have involved HCC over the last several years would not have been possible without collaborations at all scales. Regionally, our network connectivity has been more affordable due to planning and cost-sharing with partners in the Great Plains Network. This has facilitated greatly improved access to Internet2 and ultimately the world. Collaborations with OSG as well as XSEDE have facilitated access to resources but have also provided a broader context for the development and contributions of researchers at HCC. These personal contacts have been as valuable as any hardware access. Finally, the LHC is one of the largest international collaborations in history. HCC has been privileged to participate in this effort, and has recently been able to celebrate with colleagues worldwide at the first major breakthrough this project has produced. Through further collaboration, we look forward to future advances and discovery.

# On the End of Paper Based Communication

**Perry Alexander**, Director, Information and Telecommunication Technology Center (ITTC), University of Kansas

The post-PC world is upon us. On Christmas Day, 2011, the number of tablet computer owners in the world doubled. Apple consumes more silicon than any company on earth, yet it controls only 12% of the PC market. Physical media for music and video are antiques and printed media is next. I have not printed a single technical paper since purchasing my iPad a month after they came out. I have not visited a physical library in over 10 years.

We are consuming more information in more ways than ever in our history. Information is ubiquitous and always available. The industrial revolution snatched up our information infrastructure and made its products commodities. Specialized tools targeting specific types of information and consumers are replacing general purpose desktop PCs at a rapid pace. Truly, the post-PC world has arrived.

Yet…

Accounting still needs physical receipts for my trips. They will scan them and destroy them, but they need physical receipts. Wouldn't it be simpler for me to scan and send them myself? The "sign and return" two-step is still far too common - receive a document in email; print it; sign it; scan it; send it back; and throw the copy in the recycling bin. What if someone cuts a signature out of a previously signed document and "signs" the document without my knowledge? Contracts, POs, and the paperwork of business is still quite literally paper. Wouldn't it be simpler to never print them at all? Homework, projects, exams, and textbooks are still largely physical, linear, and expensive. Why is this? Whatever happened to the paperless office promised two decades ago? The technology is most clearly here and available. What's missing? What does paperless mean?

Paperless means literally less paper, lowering costs, and producing less waste. These are great things. However, there is another side to this coin that institutions are finally recognizing. What we are really doing is shifting, rather than eliminating consumption. Instead of reams of paper and stamps, we now consume terabytes of bandwidth and storage. Thus, we need more information technology. We need greater bandwidth and storage, new kinds of data archives, more software with far higher complexity, all with increased reliability and ubiquity requirements. We may need more people, or we may need less people, only time will tell. However, we certainly need *different* people or the same people with *different* skills. New skills for everyone. New protocols for information exchange. New mechanisms for decision-making. New classes of people who fix our technology

and make sure it is up-to-date and always there. Still, this will do nothing for the print-and-sign protocol because while the paper carries information we see, it also establishes trust. We need *new ways of establishing trust* that reflect our new ways of storing and transmitting information.

For a researcher who looks at high-assurance systems and security, *the need to shift trust from physical to virtual things* is fascinating. When we make things electronic, we eliminate the traditional places where we root our trust – the physicality of receipt or a contract, the ink of a signature, the weight of a book, or the authenticity of a homework submission or exam. All of these things definitely convey information and there are more efficient, more effective ways of transferring and storing that information. But trust is simply not keeping up – we need better models for establishing trust.

Certainly there is risk in paperless trust. There is also risk in physical artifacts of trust – that risk is simply obscured by familiarity. We have all heard "it's always been done this way", "I've never seen it done that way", or my personal favorite "the University of (your favorite rival here) does it this way." We all have Rasputin whispering in our ear about lawsuits, audits, FERPA, HIPAA, security and friends that surely enough paper will protect us from. But paper won't protect us. We know it won't because it never has. So, let's move forward.

A signature, a sealed envelope, and handshake are all physical things where we root our trust. When a trusted party signs a letter, the contents of that letter are trusted to come from the associated agent. When a letter is sealed in an envelope, the contents are trusted to be delivered in a confidential manner. Finally, when information comes from a trusted source through other trusted parties, the contents are trusted.

The tools of virtual trust are cryptographic functions for encrypting and signing messages and cryptographic protocols for exchanging information. Encrypting information with a key provides the same trust as a sealed envelope – no unauthorized agent may access transmitted information that is encrypted. Signing information with a key provides the same trust as a physical signature – no unauthorized agent may generate a signature over information. Protocols provide us methodologies for using encryption and signing to exchange information.

Exchanging information securely using encryption and signing incorporated into protocols will emerge as a viable solution for trust management. They key is asymmetric key encryption. In asymmetric key encryption one key is used for encrypting secrets while another is used to decrypt. If Alice and Bob want to communicate securely, they exchange their public keys and keep their private keys secret. If Alice wants to send Bob a confidential file, she simply encrypts it with is public key. Upon receipt, Bob can decrypt it with his private key. If Cindy intercepts Alice's message, without Bob's private key she cannot decrypt it. This guarantees *confidentiality* in communication.

Conversely, if Bob wants Alice to know it is he sending the file, he can create a cryptographic signature for the file. Bob creates a signature from his file using a hash function. That signature is then encrypted with Bob's private key. When Alice receives the file that Bob sent, she can decrypt the signature with his public key and check the fingerprint against the file she received. If Cindy sent the file posing as Bob, Bob's public key cannot decrypt the signature because Cindy's key created it. Thus, Alice would know that Bob did not send the file. This guarantees *integrity* in communication.

Asymmetric key cryptography gives us the tools to electronically replace envelops that guarantee confidentiality and physical signatures that guarantee integrity. *Protocols* then specify how those tools are used in practice. One such protocol example is S/MIME for sending privacy-enhanced email. S/MIME is incorporated in virtually every modern email client. It manages keys, ensures your messages are encrypted and signed, and decrypts and checks signatures on incoming email. It literally eliminates the sign and return protocol discussed earlier. There are far more sophisticated protocols for achieving a wide variety of security and privacy goals. We are seeing these protocols implemented in everything from software distribution systems to lab information maintenance.

Establishing trust electronically has its problems. Key management – particularly revocation of compromised keys – is an ongoing area of research and development. But the tools are there for us to use. The time is now for us to move forward and begin to put trust on equal footing with information in the electronic world.

# The Role of Information Systems in Clinical and Translational Research (Frontiers: The NIH Clinical and Translational Science Award Driving Information Systems)

**Paul F. Terranova[1], Richard J. Barohn[2], Lauren S. Aaronson[3], Andrew K. Godwin[4], Peter Smith[5], and L. Russ Waitman[6],** University of Kansas Medical Center

Elias Zerhouni, former NIH Director, charged the NIH with building a biomedical research enterprise that enhanced the translation of scientific discoveries into improved health care for our nation (1). One component of that charge was to establish the NIH Clinical and Translational Science Award (CTSA) program, which currently is located in the National Center for Advancing Translational Sciences (http://www.ncats.nih.gov/research/cts/ctsa/ctsa.html. The goals of the CTSA program are 1) to improve the manner by which medical research is performed in the US, 2) to move discoveries in the lab more quickly into the clinic, 3) to integrate clinical research into communities, and 4) to train the next generation of medical researchers in the art of clinical and translational research. Approximately 60 institutions in 30 states including the District of Columbia and Kansas are part of the CTSA consortium working toward these goals.

Headquartered at the University of Kansas Medical Center, the NIH-CTSA-supported Frontiers program, more formally called The Heartland Institute for Clinical and Translational Research, is a network of scientists from institutions in Kansas and the Kansas City region (www.FrontiersResearch.org). The Frontiers program is developing more efficient use and integration of biorepositories, genomic information and biomedical informatics which are important components for attaining the CTSA goals. Specific goals of Frontiers are to: 1) create a new academic home with innovative education and training programs for clinical and translational investigators that will transform the type, and increase the number of faculty and investigators needed to bring discoveries and research findings more rapidly to the point of care; 2) provide an enhanced coordinated translational research infrastructure that will speed the process from discovery to community research in Kansas and the Kansas City region; and 3) actively engage the community in developing, testing, and disseminating translational research through existing and new networks in Kansas and the Kansas City region.

[1] Vice Chancellor for Research; [2] Director, Frontiers; [3] Deputy Director, Frontiers; [4] Professor & Director, Molecular Oncology; [5] Director, Institute for Neurological Discoveries, Co-Director, KIDDRC; [6] Director, Medical Informatics.

The infrastructure needed (Frontiers goal #2) to speed the process of discovery, includes, in part, fundamental components such as bio-repository, genomics, and biomedical informatics. Each of these promotes and enhances discovery and each is foundational for moving discoveries from the bench to the bedside and becomes the foundation for this presentation. Each of these is described in order to provide an overview of how each is a component of translational research.

**Frontiers: The Heartland Institute for Clinical and Translational Research**

Many partners are involved with accomplishing the goals of our Frontiers program. These partners include academic institutions (University of Kansas Medical Center in Kansas City, University of Kansas School of Medicine in Wichita, University of Kansas at Lawrence, University of Missouri at Kansas City, and the Kansas City University of Medicine and Biosciences) and health care institutions and centers in the region (University of Kansas Hospital, Wesley Medical Center in Wichita, Via Christi Health System in Wichita, Kansas City, VA Medical Center, Children's Mercy Hospitals and Clinics, St. Luke's Health System, Truman Medical Center, Swope Health Services, Center for Behavioral Medicine, and Center for Practical Bioethics. The basic infrastructure of Frontiers includes several components described below.

1. Clinical Research Development Office. This is the literal and virtual (via a web portal) front door where investigators go to access information about services and resources. Information on the investigators research interests are logged so they may be contacted by Frontiers faculty regarding special initiatives and for collaboration. Thus, this office brings together the scientific expertise and other resources to support faculty projects, ensuring appropriate multidisciplinary contribution and inclusion of relevant components of the CTSA such as biostatistics, biomedical informatics, community engagement and ethics programs.
2. Clinical and Translational Research Education Center. This is the site of educational and career development programs that use teaching and learning technologies and faculty/staff to educate students and faculty at many levels of development. This Center provides innovative programs for mentoring skills and development, formal training in the entrepreneurial skills necessary to bring new discoveries to the market, and a program for minority recruitment and retention. This Center also includes a NIH TL1 pre-doctoral program and a NIH KL2 post-doctoral program. In addition to formal training in research methods, scholars and trainees take advantage of a mentorship program and become integrated into multidisciplinary research teams.
3. Biomedical Informatics. This program provides information resources, services, and communication technologies in support of translational research. Russ Waitman, PhD, associate professor and director of Medical Informatics in the Depart-

ment of Biostatistics leads this program and is providing a chapter on this topic in this Merrill conference (2012). The program has several goals: a) to provide a portal for investigators to access clinical and translational research resources, track usage and outcomes, and provide informatics consultation services; b) to create a repository to integrate clinical and biomedical data for translational research aligned with national standards; c) to advance medical innovation by linking biological tissues to clinical phenotype and research laboratory results for phase I and II clinical trials; and d) to leverage an active, engaged statewide telemedicine and Health Information Exchange to enable community based translational research.

4. Biostatistics. This program, offered through the Department of Biostatistics, provides the statistical and data management support for Frontiers studies.
5. Clinical and Translational Science Unit (CTSU). This unit is a new state-of-the-art facility with capability to conduct first-in-human proof of concept trials as well as traditional clinical research activities within the unit and in clinics and the community (using our "CTSU without Walls"). This program also developed a clinical research participant identification and registration system that our Biomedical Informatics program is further developing to provide easy cohort identification and contact information on patients willing to consider participating in research and to incorporate national developments in this area.
6. The Institute for Advancing Medical Innovations. This program provides infrastructure for drug and device development that lead to clinical applications.
7. Translational Technologies Resource Center. This program provides access to specific technologies through four cores: In vivo Imaging, Cell and Tissue Imaging, Molecular Biomarkers, and a Biological Tissue Repository. The overall goal of this program is to eliminate barriers to incorporating new technologies and research approaches into research studies for both new and seasoned investigators by providing access to equipment and expert faculty collaborators in these technologies.
8. Pharmacokinetics/Pharmacodynamics (PK/PD). This program provides PK/PD support to investigators who prefer to focus on other aspects of their investigator-initiated trials. The program also supports clinical and translational research by training investigators who wish to develop such skills to become competent in designing, analyzing and interpreting PK/PD data obtained from their investigator-initiated trials, and by providing a training ground for graduate and post-doctoral students on PK/PD.
9. Personalized Medicine and Outcomes Center. This program focuses on T2 research innovations and support and has five cores: Analysis of Large Databases; Quality Assessment/Quality Improvement Analysis;

Economics, Health Status & Decision Analysis; Survey Design and Qualitative Research Support; and Outcomes Education.

10. Pilot and Collaborative Studies Funding. This program serves as the centralized Frontiers resource and catalyst for soliciting, reviewing, awarding and tracking clinical and translational research pilot and feasibility studies.
11. Community Partnership for Health. This program supports and assists investigators with accessing diverse populations for participation in clinical and community based research projects—historically one of the larger barriers for many clinical investigators. It ensures investigators are skilled and knowledgeable about working in and with diverse communities through a novel training and certification program on community engagement for research. This program is provided online and will be coordinated with all other annual investigator certifications (e.g., on Human Subjects and Biosafety) required for maintaining Institutional Review Board approvals.
12. Regulatory Knowledge and Support. This program supports investigators to ensure their clinical and translational research is in compliance with ethical and regulatory standards.
13. Ethics. This program provides individual ethics consultations to Frontiers investigators and formal research ethics education and training programs.

**Bio-repository**

The bio-repository includes numerous components requiring integration of multiple sources of information flow with a goal of enhancing discovery to improve health. The bio-sample repository has several major functions including recruitment, sample acquisition, processing, storage, distribution and analysis. The repository also includes tissues, bodily fluids and cell lines. The process usually begins with recruitment and informed consent of the patient in order to collect bodily tissue and/or fluids (e.g., serum or plasma). Informed consent may occur during pre-admission testing, an outpatient visit or upon admission to the hospital. Sample collection may occur in the hospital, operating room and/or outpatient clinic. Samples are proactively acquired for future testing and may eventually be sent to multiple recipients. Information on personal and family histories, clinical intervention, and lifestyle factors is collected for each sample as well as: demographic, family history, medical history, epidemiologic risk factors, and clinical history. Annotated, high quality samples are collected and stored. Specific procedures are followed for the tissue banking scheme. For example, surgical specimens are collected by the pathologist with the aid of their physician assistant. Select tissue is identified for sampling and this information is entered into a database. A sample is then snap frozen and kept in liquid nitrogen or prepared for frozen sectioning. Samples frozen in liquid nitrogen may be stored and/or prepared for DNA and RNA extraction. Another sample of the same tissue is placed in fixative for sub-

sequent histopathological analysis by storing in paraffin blocks for subsequent preparation of histological sections including routing staining, immunostaining, or tissue microarrays. Blood is separated into serum or plasma, red blood cells, leukocytes, DNA and stored at -80C. Blood samples are banked within 4 hours of draw after being barcoded. Sample quality is critical to the success of the bio-repository. Quality measures taken include compliance with best practices, qualification for membership in the International Society for Biological and Environmental Repositories, insisting on standard operating procedures, and certification by the College of American Pathologists (and others). NIH supported projects and others will/may require appropriate mechanisms for data sharing. Thus, patient consent, sample processing and inventory, questionnaire data, medical records and tumor registry data (when appropriate) must all be stored in databases, managed, and appropriate samples identified when requests are made for procurement. Bio-repository management is complex with many components requiring oversight such as equipment, supplies, dedicated resources, liability, data management, standard policies and procedures, quality control, and compliance. Access to bio-samples is regulated by an advisory committee which considers requests from within and outside of the medical center.

Our goal is to enter genome and other types of 'omic' data such as metabolome, transcriptome and proteome into the system. Thus, sequencing data, genotype, methylation, expression, single nucleotide polymorphism data, chromatin remodeling, etc. are collected and entered. The samples are annotated and the data placed in the Healthcare Enterprise Repository for Ontological Narration (HERON—see Waitman presentation in this Merrill series, 2012). HERON also integrates with billing records, medical records, and the KU Hospital Tumor Registry along with all of the 'omic' and other types of data. The Bio-sample Repository supports several programs including, Frontiers, Early Detection Research Network, Alzheimer's Disease Center, Clinical Trials and Protocol Support Lab, the Consortium of Investigators of Modifiers of BRCA1/2, Pharmaceutical and Biotech collaborations, the Triple Negative Breast Cancer Network, faculty collaborations including investigator initiated studies, and the Gynecological Oncology working group.

Compliance is important in the bio-repository and includes federal and local components such as Health Insurance Portability and Accountability Act and Institutional Review Board of protocols.

**Bioinformatics Services**

Bioinformatics services (http://www2.kumc.edu/siddrc/bioinformatics) at KUMC are provided largely by the Smith Intellectual and Developmental Disabilities Research Center (SIDDRC; http://www2.kumc.edu/siddrc/) which is a collaborative effort with KU at Lawrence (http://kiddrc.kumc.edu/). The Center is supported by a NIH grant (HD002528), which has been held for 47 consecutive years and the grant will continue at least into its 50th year. Additional support for bioinformatics comes from NIH including the Center for Biomedical

Research Excellence in Cell and Developmental Biology (RR024214) and the Kansas Idea Network of Biomedical Research Excellence (GM103418, RR016475). Bioinformatics studies and utilizes methods for storing, retrieving and analyzing biological data, such as DNA, RNA, proteins and their sequence as well as their structure, function, pathways and interactions. The overall mission of the core facility is to advance the understanding of integrative functions in biological systems, including human, through the application of computational models and data analysis with focus on Next Generation Sequencing Data Analysis and Microarray Data Analysis. The Bioinformatics core provides consulting, data analysis solutions and analysis of large-scale biological datasets produced by high-throughput genomics experiments. Thus, the core identifies opportunities and implements solutions for managing, visualizing, analyzing, and interpreting genomic data, including studies of gene expression (RNA-seq and microarrays), pathway analysis, protein-DNA binding (e.g. ChIP-seq), DNA methylation, and DNA variation, using high-throughput platforms in both human and model organisms.

A variety of services in bioinformatics and computational biology are provided by the Bioinformatics Core. For example:

**High-throughput sequencing**

- **RNA-Seq:** provides an unbiased deep coverage and base level resolution of the whole transcriptome.
- **Chip-Seq:** combines chromatin immunoprecipitation with high-throughput sequencing to provide an unbiased whole genome mapping of the binding sites of DNA-associated proteins.
- **Whole Genome Sequencing:** sequences the whole DNA sequence of an organism's genome.
- **De novo Sequencing:** provides the primary genetic sequence of an organism.
- **Metagenomic Sequencing:** sequencing and identifying the genomes of whole microbial communities.
- **Methyl-Seq:** analysis of methylation patterns on a genome wide scale.

**Microarray analysis**

- **Affymetrix 3′ Expression Arrays:** target the 3′ end of genes.
- **Affymetrix Exon Arrays:** provides expression levels for every known exon in the genome.
- **Affymetrix miRMA Arrays:** provides measurements of small non-coding RNA transcripts involved in gene regulation.
- **Affymetrix Genome-Wide Human SNP Array:** copy number analysis
- **Affymetrix GeneChip Tiling Arrays:** gene regulation analysis

**Biological Functional and Pathway Analysis:** software from Ingenuity Systems analyzes expression data to ascertain the top biological functions and pathways associated with them. **Biological Literature Survey:** software from Acumenta (Literature Lab) that performs data mining tasks on experimentally derived gene lists. **miRNA Target Prediction:** in house software and open source software such as TargetScan, miRanda for detecting genomic targets for miRNAs.

**Transcription Factor Binding Site Prediction:** in house software and open source software such as MEME, Homer, PGMotifScan to identify protein DNA interaction sites.

**Medical Informatics**

The KUMC Medical informatics program is discussed in more detail in another chapter of this Merrill document (2012). The work of the medical informatics division (within the Department of Biostatistics in the School of Medicine) includes the following:

- HERON (http://informatics.kumc.edu/work/wiki/HERON) healthcare enterprise repository
- CRIS (http://biostatistics.kumc.edu/bio_cris.shtml) Comprehensive Research Information System powered by Velos
- REDCap (http://informatics.kumc.edu/work/search?q=redcap) for REDCap at KUMC and (http://project-redcap.org/) for REDCap at the CTSA consortium.
- RequestTracking (http://informatics.kumc.edu/work/wiki/RequestTracking) place for notes on using REDCap for request tracking (CTSA, ADC, HERON)
- EnterpriseAnalytics (http://www2.kumc.edu/aa/ir/)
  - TEAL data warehouse (http://informatics.kumc.edu/work/wiki/TEAL)
- NetworkInnovation (http://informatics.kumc.edu/work/wiki/NetworkInnovation)
- Other informatics efforts at the University of Kansas not mentioned in the article above include:
  - The Center for Health Informatics (http://www2.kumc.edu/healthinformatics/)
  - The Center for Bioinformatics (http://www.bioinformatics.ku.edu)
  - The K-INBRE Bioinformatics Core (http://www.kumc.edu/kinbre/bioinformatics.html)
  - Bioinformatics Computational Biology Laboratory (http://www.ittc.ku.edu/bioinformatics/)
  - Biodiversity Informatics (https://journals.ku.edu/index.php/jbi)

**Reference:**

Zerhouni EA Translational and clinical science–time for a new vision. NEJM 2005, 353:1621-23.

# Research, Data, and Administration Management in the Animal Health/One Health Era

**James A. Guikema**, Associate Vice President, Office of Research and Sponsored Programs, Kansas State University

The theme of the current Merrill Conference – information systems as infrastructural priorities for university research both now and in the future – unites the research office and the information technology (IT) office at each university institution within our four-state region. There is an unprecedented demand on our IT infrastructure from nearly all sectors of our collective clientele. Students are sophisticated users of IT networks, and push the limits of what we can provide in communication, academic course content, and social networking. This is amplified when considering the needs of the distance-education arena. Our research-faculty investigators are developing larger databases that challenge our ability to archive, manage, manipulate, mine, and share essential information.

High on the list of research priorities are the 'omics' – genomics, proteomics, metabolomics, lipidomics, etc. – with the i5K genome project [sequencing and annotating 5,000 insect genomes], described elsewhere in this volume, as an excellent example. Policy makers, and the managers of funding programs at both the state and federal levels, are sending our universities mixed messages regarding what data may (or must) be shared and what data must be secured.

Relative to IT demands, our universities are looking into the tunnel at the light – not knowing if that light is in fact the end of the tunnel, or if it is the headlight of the oncoming train. One thing is certain: a sense of a deadline approaching. This was perhaps best described by Dr. Samuel Johnson: "Nothing so concentrates the mind as the sight of the gallows," and the IT challenges are relevant across all disciplines on our campuses. Each of our institutions has risen to the IT challenge in various areas and disciplines. This article seeks to place the IT / research infrastructure challenges within a 'comparative medicine' context, since these challenges touch upon research strengths of the region, of strengths within each institution, and of the growing regional opportunity represented by the relocation of a major federal comparative medicine laboratory [NBAF] to Manhattan, KS.

<u>One Health / One Medicine:</u> A 'comparative medicine' approach to cellular and molecular biology, as recognized within the United States, is an understanding that disease conditions, and the mechanisms which cause them, are similar within animal and human physiological systems. Thus, knowledge of one will enhance the understanding of the other. Creative research in either animal or human disease physiology will have a payoff in both sectors, implying common dis-

ease mechanisms, common disease interventions, and common cures.

Although largely embraced by the academic community, implying a close linkage between colleges of medicine, veterinary medicine, and basic molecular biology, the comparative medicine viewpoint is often not recognized by the agencies which fund research. For example, the National Institutes of Health [NIH] demands a firm human health rationale for funding, and recent awards providing funding for the translation of research into the health enterprise do not consider animal health. Likewise, the National Science Foundation [NSF] will rarely fund a project with a human health rationale even at the basic molecular science level.

The comparative medicine approach is more widely embraced within agencies outside of the United States, with the addition of plant science to the human/animal/plant physiology and disease continuum. An example is the Australian National Biosecurity Flagship Program, which recognizes and funds infectious disease research on a broad level. Plant innate immunity protein homologs may play important roles in the physiology of Crohn's disease and the transfer mechanisms by which bacterial DNA is inserted into host cells were first documented in plant systems. Further, hunger is a looming political issue, with diseases such as wheat stem rust UG-99 and wheat blast taking center stage. In the past few months, there has been a 30% increase in the price of bread in Iran – a staple food in that country. Thus, the broader approach to comparative medicine recognizes food and water security as legitimate areas of research concern and funding.

Special Data Management Concerns in Research: Data management issues in research have resulted from the increasingly complex data sets routinely generated which are costly to archive, to mine, and to share. Our investigators recognize that, in collaborations which span institutions, making these data sets available to colleagues can create data-streaming challenges and require compatible assessment software on each side of the data-stream connection.

Federal funding agencies differ in their approaches to data sharing strategies. Two of the largest, NSF and NIH, have taken the strategy that, if data are to be useful for the good of society, they must be widely shared. NIH requires that all publications resulting from their funding be archived in PubMed. NSF has taken one step further, and requires the submission of a data management plan with every proposal for research funding. Kansas State University has responded to these requirements by providing a mechanism for submission to PubMed and by preparing for our investigators a data management plan which is suitable to NSF. The management plan varies depending upon 1] the size of the data sets, and 2] the frequency by which the research community will access them.

The Department of Defense, however, has taken a different approach and has proposed rules which will limit the sharing of DoD funded research. They are seeking limits on the rights of investigators to publish their results, and servers which house DoD-funded datasets may be required to be isolated with access controlled. If select agents are purchased with DoD funding, it is likely that expensive

personnel surety measures must be in place for laboratory staff.

Research management also implies a concern for the welfare of the research environment. Welfare of the animal subjects is managed by the Institutional Animal Care and Use Committee established at each of our institutions. Welfare of the research staff member within the infectious disease research arena is a bit more complex, especially when the research involves pathogens considered by either the Centers for Disease Control (CDC) or the United States Department of Agriculture (USDA) to be a virulence threat to the investigator or a threat to the environment outside of the laboratory.

These two organizations have established physical specifications to meet laboratory safety goals, and these physical laboratory specifications are termed 'Biological Safety Levels (BSL)' defined as follows:

- BSL-1 – research involving pathogens which are not hazardous to investigators and which are not harmful to the environment if released.
- BSL-2 – research involving pathogens which may impact investigators in a non-serious way.
- BSL-3 – research with pathogens which may cause serious harm, but for which disease countermeasures exist.
- BSL-4 – research with pathogens for which no disease countermeasures exist.
- Facilities at the latter two levels require inspection by either CDC or USDA [or both], and obtaining permits for some work can take months or years.

Special Opportunities for the Four State Region: The four-state region represented by this conference is particularly strong in comparative medicine research activity. The greater Kansas City region, for example, is a hub of animal health economic activity, and has been dubbed the 'Animal Health Corridor'. Each of our research universities has strengths in the continuum from plant/animal/human disease physiology and has taken great strides to translate our research strengths into disease countermeasures useful in the wheat field, the cattle feedlot, or in the human hospital.

Our special strengths have also resulted in success in obtaining specialized facilities. The University of Missouri, for example, was successful in competing for a Regional Biocontainment Laboratory though an NIH competition. Ames, Iowa, is home to a USDA facility dedicated to animal health research, and the USDA Arthropod Borne Animal Disease Unit is located in Manhattan, Kansas. The State of Kansas has invested in the construction of the Biosecurity Research Institute, a BSL-3 laboratory facility associated with Kansas State University.

One tremendous opportunity looms on the horizon: the Department of Homeland Security plans to cease operations at the Plum Island Animal Disease Research Center located off the coast of Cape Cod, New York. In its place, DHS has planned the National Bio and Agro Defense Facility [NBAF] in Manhattan, Kansas, on land contiguous to Kansas State University. This facility will have both BSL-3 and BSL-4 laboratory capa-

bilities, and is designed to protect the agricultural economic sector within the United States from foreign animal diseases.

Adding NBAF to the armamentarium of infectious disease research already within our four state region will provide us with facilities unmatched throughout the United States.

# Trends in Technology-Enabled Research and Discovery

**Gary K. Allen**, CIO, University of Missouri-Columbia; Vice President for IT, University of Missouri System

**James Davis**, Vice Provost for IT and CIO, Iowa State University

It is frequently observed that technology advances at a rapid rate, perhaps following *Moore's Law* by doubling in capacity every 18-24 months. Information technology service providers, including technology manufacturers and higher education institutions, are leveraging these gains to create new types of services that have a direct and supportive impact on research computing. The IT delivery model is also changing. It's becoming more common for universities to collaborate with peers to develop services that provide an economy of scale and the functionality that meets the specific needs of higher education. Also changing is the way that university research is structured, with increased emphasis on large cross-cutting (and often multi-institutional) thrusts. The concurrent changes in technology and the way it's used leads to a dynamic environment. Higher education technology providers must become agile and collaborative to align with the campus and global research communities. In the remainder of the paper, we'll briefly explore how these emerging trends in research and information technology are giving rise to new opportunities to accelerate campus research programs.

The pursuit to create and share knowledge is the essence of the academy. Discovery, innovation, learning, and engagement are core to the mission of the university, and support for those objectives pervades all processes.

An effective information technology infrastructure (or *cyberinfrastructure*) is central to the success of a robust research program. The infrastructure must provide anywhere, anytime access to information, peer collaborators, systems, and services needed to advance the program. When cyberinfrastructure is congruent with the needs of the research program, the ensuing interdependence creates significant synergy that acts as a "force multiplier" to propel research activities forward. Although beyond the scope of this paper, we note that aligning IT with the missions of the university has a similar beneficial effect on technology-enabled learning, the ability to base decisions in institutional information through multiple analytics, and efficient engagement of broad constituent groups. It also contributes to a forward looking environment that assists with recruiting and retaining students and faculty. In a very real way, campus information technology is a strategic asset of the university.

Building an effective cyberinfrastructure to support discovery and innovation is a national priority. Campus technology and research visionaries

have been working with the National Science Foundation (NSF) Office of Cyberinfrastructure the past few years to develop strategies for advancing campus research in a comprehensive and coordinated way. The National Science Foundation Advisory Committee for Cyberinfrastructure Task Force on Campus Bridging offered insightful recommendations for research universities and federal funding agencies in their 2011 final report[1]. One recommendation calls for a "healthy national cyberinfrastructure ecosystem" and a rethinking of some of the barriers to large-scale collaboration. We end this paper with a call for a regional approach to nurturing the campus dialog that brings national cyberinfrastructure efforts to researchers in a consequential way.

## Two Research Trends Impacting Cyberinfrastructure

### "Big Data Science"

The escalating capacity and reduced cost across almost all components of research cyberinfrastructure is encouraging the expansion of computational models to yield results with improved fidelity. As the problem size scales up, so does the demand for computing resources. Large-scale high performance computing clusters, once used by just a handful of disciplines for specialized problems, are now an essential tool in any area where timely results are important. Large computational problems nearly always consume and produce significant collections of unstructured data that must be stored and transmitted. Additionally, several federal funding agencies now require that project data be retained and made available to the public so that the reported results can be validated and used for follow-on research (for example, see National Science Foundation data management plan guidelines[2]). Even with the rapidly increasing capacity of contemporary storage, the management of large collections of data is demanding. New storage architectures have been developed to implement hierarchies with integrated networking to balance performance and cost, yet it remains challenging to perform even basic operations such as searching large data stores or archiving collections. The complexity of maintaining large data stores coupled with curation requirements and rapidly expanding security requirements (e.g., FISMA compliance[3]) makes a compelling case for developing an institutional approach to data management.

Another challenging component of processing large research data sets is simply moving them quickly and reliably, for example between a laboratory instrument and a HPC cluster. Even with contemporary 10-gigabit or the emerging 100-gigabit network connections, data transfers can often be measured in hours. This is exacerbated when large data sets are moved between institutions, sometimes at "Internet speed". Additional barriers pertaining to format and data compatibility can arise.

### Interinstitutional Research Collaboration

As we recognize the need to substantially improve the capacity of research cyberinfrastructure, we also note that a slow shift in the way scientists collaborate brings new requirements to create "virtual communities" for partici-

pants to gather (virtually) to share ideas, information, systems, and services. While it was more common in the past for grants to have one or two investigators, there has been a steady growth in large multidisciplinary grants with teams that span multiple universities, centers, and corporations. In one approximation for the degree of interinstitutional collaboration, a 2012 National Science Foundation survey of R&D expenditures at universities and colleges[4] noted that the "pass-through" funds represented 7% of the total academic R&D expenditures in FY 2009 as compared to 5% in FY 2000. Project teams expect that technology will mitigate the impact of distance and create a community where participants can interact as if they were located in the same space.

These two aspects taken together have implications for research cyberinfrastructure:

- An increased demand for HPC resources in disciplines that have not in the past been notable consumers of those services.
- A pressing need for significantly larger compute clusters capable of scaling up problems to enable new discoveries.
- The need for specialized one-off computing resources such as GPU-based HPC clusters.
- The ability to transmit, store, and share massive data collections, while addressing cost, security, curation, and backup. New business models must be developed to support the management of data beyond the duration of a research grant, in partnership with institutional libraries.
- Seamless support for collaboration, including video conferencing and shared data, systems, and documents. Services need to be layered on a commonly used identity management paradigm that supports federated identities, such as the Internet2 InCommon suite.

Implicit too is establishing a collaborative relationship between campus technology providers, institutional leadership for research, and research centers and their faculty and staff. As previously noted, the case for providing an enabling research technology infrastructure is clear, but the issues surrounding complexity and cost require an institutional approach to obtain an effective outcome, given the rapid changes in technology and the evolving way that campus research is carried out.

**Information Technology Trends Impacting Research Cyberinfrastructure**

Market forces have created demand for a new operational model within IT structures. These forces include the economy, price of utilities, personnel costs and IT complexity, among others. Organizations can no longer afford to manage everything on their own. IT is in the largest outsourcing trend in history, and public cloud Infrastructure as a Service (IaaS) will be a key component of outsourcing decisions because it offers commoditized infrastructure and increased agility. This allows IT to focus on core business outcomes. In reaction to these forces, public cloud IaaS providers have developed services that aim to solve many of the business issues IT organizations face[5].

In higher education institutions, these trends are clearly evident in areas of administrative computing and academic technology services supporting teaching and learning. In these areas, migrations toward outsourced or shared solutions are well underway. A different rate of evolution is evident in IT services supporting research activities; these frequently require architectures and applications that are not "commoditized" in the sense that the research goals are sometimes optimally met using non-standard IT toolkits not readily available in the marketplace. Historically, there has not been a sufficient market demand for some of the commonly needed research IT workflows to merit investment by for-profit providers. Additionally, the information security and data management issues at play with research-related IT often requires approaches far different from other types of information systems from research universities. This includes issues such as intellectual property protection (e.g., for investigators, or corporate sponsors), data classification (e.g., for certain types of federally funded research), and data curation (e.g., for mandated public dissemination of research results).

Given these additional concerns and demands for management of research IT infrastructure and data, new models are rapidly developing among communities of use whose members have these requirements in common. Research consortia for IT infrastructure and services are now developing in ways reminiscent of similar approaches used by academic libraries developed and deployed over the past several decades. Significant cost and complexity drivers for these research tools add significant momentum for this evolution, along with the heightening of interdisciplinary research approaches that blend data systems and tools that provide new perspectives into research questions.

Joint efforts, many of which are regionally focused, are also growing in number as institutions work together to develop new approaches to satisfy their research cyberinfrastructure needs. Examples include: the Renaissance Computing Institute (RENCI)[6], Rocky Mountain Supercomputing Center (RMSC)[7], University of Chicago Computing Institute (UCCI)[8], Computational Center for Nanotechnology Innovations (CCNI)[9], and the Victoria Life Sciences Computation Initiative (VLSCI)[10]. In many cases, these efforts include both research universities and corporate partners. They are frequently specifically targeted at not only enhancing research capacity, but also providing nuclei of cyberinfrastructure equipment and expertise designed to foster economic development efforts.

Federal agencies and higher education organizations such as Internet2 are directly supporting these approaches. The National Science Foundation has long been an active supporter of cyberinfrastructure enhancement efforts across the U.S. The Global Environment for Network Innovation (GENI)[11] Project was established in 2007 by the NSF. The goal was to provide an environment for the development of novel networking approaches and applications that obviate the constraints of the current internet and wireless network environment such as might be observed when trying to

move very large datasets at very high speeds. As indicated above, the research computing needs of higher education increasingly demand the ability to move very large volumes of data at very high speeds; such ability is limited on today's established networks. Accordingly, GENI targets promoting networking innovations that utilize existing network infrastructure in ways that are increasingly scalable, flexible (in terms of routing and quality of service), and secure. Significant efforts by GENI program participants relate to software and networking protocols supporting "software-defined networking" as exemplified by OpenFlow. OpenFlow creates virtual network "slices" that operate on existing network infrastructure and share that infrastructure with other networking protocols. Numerous universities are experimenting with the development and deployment of OpenFlow, and both Internet2 and National LambdaRail are implementing OpenFlow in their networks. In 2011, the Open Networking Foundation (ONF)[12] was created as a not-for-profit entity focused on further development and standardization of software-defined networking. ONF is currently supported by more than 70 global companies, and is accelerating the evolution of these new approaches to networking.

The recently announced Internet2 Net+ services and Innovation Platform initiatives are providing both middleware and end-user services targeted specifically at higher education research activities, offered in ways that are congruent with policy frameworks required by universities, and that meet regulatory and compliance concerns. These offerings include extremely high-bandwidth network connectivity, dynamic configuration of networking protocols, data security, and a new generation of monitoring and management tools. For example, Internet2's Advanced Layer 2 Service is a reliable Layer 2 transport environment that provides a flexible end-to-end, high-bandwidth and deeply programmable environment. The Advanced Layer 2 Service builds on Internet2's NDDI/OS3E initiative to provide network connections that can be used to easily create VLANs with a range of characteristics, with reachability through the network, regional networks, campuses, and partners like ESSnet, GEANT and other inter-domain enabled global networks. (see the Internet2 Innovation Platform[13]).

**Call to Action**

Discovery and innovation are core to the academy mission. Information technology can enable remarkable research when there is alignment between the IT resources available (the cyber-infrastructure) and the needs of the research enterprise. Alignment is challenged by rapidly changing trends in the information technology field, by dramatically increasing computational needs, and also by a shift toward large interinstitutional grants where collaborators work at a distance. In recognition of the need for institutions to collaborate to achieve the scale required to address these issues, several national and international initiatives have formed around the goal of improving campus cyber-infrastructure.

One of the most challenging aspects of determining what would work best locally is engaging campus leadership and key research faculty in explorations of adapting consortia services and experiences to a local context. Given the long standing collaboration among the "four corners" universities at the Merrill workshops, we believe that there's an important role for a regional approach to developing strategies to bridge between the campus and national research cyberinfrastructure initiatives. Additionally, all of the represented states in the Merrill workshops are EPSCoR-eligible; this could represent a significant opportunity to obtain federal funding support (e.g., in the form of NSF EPSCoR Track 2 grants) to begin creating a regional support model for cyberinfrastructure. It is proposed that follow-on meetings with the chief research officers, chief academic officers, and chief information officers of the universities that participate in the Merrill workshop explore this further.

**References**


1. National Science Committee Advisory Committee for Cyberinfrastructure Task Force on Campus Bridging. Final Report. March 2011. Available from: http://www.nsf.gov/od/oci/taskforces/TaskForceReport_CampusBridging.pdf
2. National Science Foundation data management plans. www.nsf.gov/eng/general/dmp.jsp
3. National Institute of Standards and Technology Federal Information Security Management Act (FISMA). http://csrc.nist.gov/groups/SMA/fisma/index.html
4. Hale, Katherine. "Collaboration in Academic R&D: A Decade of Growth in Pass-Through Funding" NSF 15-325, August 2012. http://www.nsf.gov/statistics/infbrief/nsf12325/
5. Hilgendorf, Kyle. Market Profile: Public Cloud IaaS, Gartner, June 12, 2012
6. Renaissance Computing Institute (RENCI). http://www.renci.org
7. Rocky Mountain Supercomputing Center (RSMC). http://www.rmscinc.org
8. University of Chicago Computing Institute (UCCI). http://www.ci.uchicago.edu
9. Computational Center for Nanotechnology Innovations (CCNI). http://www.rpi.edu/research/ccni
10. Victoria Life Sciences Computation Initiative (VLSCI). http://www.vlsci.org.au
11. Grochow, Jerrold M. "GENI Workshop on Layer 2/SDN Campus Deployment: A Report of the Meeting (July 2011)". Retrieved 9/4/2012 from http://www.educause.edu/library/resources/geni-workshop-layer-2sdn-campus-deployment-report-meeting-july-2011
12. Open Networking Foundation (ONF). https://opennetworking.org
13. Internet2 Innovation Platform FAQ. http://www.internet2.edu/pubs/Internet2-Innovation-Platform-FAQ.pdf

# Communicating Science in an International Arena: The i5k Initiative

**Susan J. Brown**, University Distinguished Professor of Biology,
Kansas State University

A few years ago I described to this group an international collaboration that was organized to analyze the genome sequence of the red flour beetle *Tribolium castaneum.* This small beetle, a cosmopolitan pest of stored grain, is a premier model organism for studies in developmental biology and pest management. When the Tribolium genome was sequenced, the cost of sequencing an insect genome of approximately 200 Megabases (10 fold smaller than the human genome) still exceeded 2-3 million dollars. The project received support from the National Human Genome Research Institute (NHGRI) and the USDA. Since then, a new generation of sequencing technology has been introduced.

As technical advances lower the costs of high through-put sequencing, genome sequencing is no longer limited to large sequencing centers and external sources of funding. Individual academic and industrial research groups are getting involved, sequencing genomes to address questions in biology, physiology, biochemistry and phylogenetics. This avalanche of data creates special challenges in how to store, share and analyze huge extremely large datasets with an international cohort of collaborators.

As sequencing costs plummet, sequencing projects are directed toward more ambitious goals. The G10K proposes to sequence 10,000 vertebrate animals to address questions of animal biology and evolution[1]. The 1000 Genomes Project will characterize variation in the human genome to understand the roles such variation has played in our history, evolution and disease[2]. The goal of the 1KITE project (1000 Insect Transcriptome Evolution, http: //www.1kite.org/) is to construct a more complete and accurate phylogenetic tree of insects, based on gene sequences. At K-State, we established a center for genomic studies of arthropods affecting plant, animal and human health. This center established an annual symposium focused on arthropod genomics that has organized and energized the community of arthropod genomic researchers.

In March 2011, the i5k initiative was announced in a letter to the editor of the journal Science[3]. This initiative is based on the understanding that these new sequencing technologies allow us as a research community to sequence not only all the species of interests, as in the G10K, but we can also sequence all the individuals of interest, as in the 1000 Genomes Project. Given the goal of the i5k to sequence the genomes of 5000 insect and related arthropod species, the first challenge is to determine which

species to sequence, and how to prioritize them. Moreover, it is important to justify not only each species, but the entire i5k project. As stated in the Science editorial, the i5k consortium is interested in all insects of importance to agriculture, medicine, energy production, and those that serve as models, as well as those of importance to constructing the Arthropod Tree of Life[4].

Sequencing 5000 insect genomes will provide a wealth of information about the largest group of species on earth. First, the genome sequences will serve as a resource for gene discovery. Identifying genes of interest is the first step in understanding mechanisms of pesticide resistance or disease. Genome sequences will also serve as substrates for analysis, to detect signatures of selection and structural rearrangements associated with their evolutionary history. Genome sequences will also serve as gateways for biological inquiry, providing a "part lists" of genes for investigation.

Arthropods display a breadth of biodiversity unparalleled in other animal groups. In approaching the ambitious goal of sequencing 5000 insect genomes, the i5k initiative began by taking suggestions from insect biologists and prioritizing them based on several criteria. First, the scientific impact of sequencing the genome of each candidate species is considered. A high priority candidate might be relevant to problems in agriculture, human health, evolution or ecology. Some species are used as model systems for basic biological studies, while others are serious pests or vectors of diseases. Still others fulfill beneficial functions, including pollination or biological control of other pests. Second, although access to genome sequences is likely to attract the interest additional researchers, the size of the research group that is already focused on a particular insect or group of insects is considered.

Additional prioritization criteria address the feasibility of a genome sequencing project. Genome sizes vary widely among arthropods and those with relatively small genomes (100-500 Million bases) are usually given higher priority. In general, smaller genomes, containing less repetitive DNA , are easier to sequence and assemble. This explains why genomes of viruses and bacteria were the first genomes to be tackled, and continue to proliferate apace in genome databases (e.g. the genome database at the National Center for Biomedical Information, NCBI). Sample availability is also considered. Model organisms, which have been reared in the lab for several years are often inbred or can be inbred, reducing heterozygousity, which also confounds sequence assembly. In the absence of an inbred strain, a sample from a single large specimen, or haploid organisms, such as hymenopteran males are given higher priority. As technology advances and the costs of sequencing continue to decline, these issues may be of less significance. But regardless of the amount of data that can be afforded, the assembly algorithms must also improve.

The next generation sequencing technologies that have yielded the most dramatic cost reductions produce exceedingly large datasets. Many algorithms have been developed to assembly

a genome from these whole genome shotgun reads, and all are computational expensive. In addition, the resulting assembly must be assessed for accuracy and coverage. Once a satisfactory assembly is a produced, it must be annotated, archived and made available to the research community. After the initial, mostly automated, annotation, a new genome sequencing project is described in a publication which marks the "release" of the genome and the genome project is simultaneously submitted to NCBI; the raw reads are submitted to the "short read archive" and the entire project is submitted as a bioproject (http://www.ncbi.nlm.nih.gov/bioproject).

At K-State, we are using NGS sequence data to improve the original Tribolium castaneum genome assembly. In addition, we have sequence the genomes of three related species. Each project generates terabytes of data to be archived, analyzed, and shared with an international group of collaborators. To update the original T. castaneum genome assembly we used short reads supplied by our colleagues in Germany, which were hand delivered at an international AGS symposium on seven CDs. We worked with the DNA sequencing center at the University of Missouri to generate the sequence data for the related species. Just downloading these datasets took several hours each. Although the initial assembly, annotation and analysis take a team of bioinformaticists and biologist several months, the end product is not static. Each project is dynamic; a web portal for community feedback is an essential component. To be useful, genome annotations must be periodically updated by a combination of community feedback and continued analysis. We currently provide community access to three genome projects at K-State through agripestbase (agripestbase.org). To solve problems in distributed resource sharing we are working with our high performance computing (HPC) group to explore solutions offered by Globus[5].

While the community of scientists associated the i5k continues to expand, it is important to point out that several are right here at Universities that participate in the Merrill retreat. These include Amy Toth at Iowa State University, Chris Elsik at the University of Missouri and Jennifer Brisson at the University of Nebraska, Lincoln as well as a host of colleagues at K-State.

Although each arthropod genome that is sequenced provides a wealth of new data, there is a broader research perspective to consider. Each species is part of a broader community including parasites, predators, prey, competitors, endosymbionts, and conspecifics. Sequencing the genomes of the interacting members of these communities will provide the foundation for arthropod community genomics and even larger datasets to consider.

**References Cited:**

1. (2009). Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. J Hered *100*, 659-674. 19892720
2. Consortium, G.P. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061-1073. 20981092
3. Robinson, G.E., Hackett, K.J., Purcell-Miramontes, M., Brown, S.J., Evans, J.D., Goldsmith, M.R., Lawson, D., Okamuro, J., Robertson, H.M., and Schneider, D.J. (2011). Creating a buzz about insect genomes. Science *331*, 1386. 21415334
4. Maddison, D.R., Schulz, K.-S., and Maddison, W. (2007). The Tree of Life Web Project. Zootaxa *1668*, 19-40.
5. Foster, I., Kesselman, C., NIck, J., and Tuecke, S. (2003). The Physiology of the Grid. In Grid Computing: Making the Global Infrastructure a Reality, F. Berman, G. Fox and T. Hey, eds. (Chichester, UK: John Wiley and Sons, Ltd).

# Advancing Clinical and Transformational Research with Informatics at the University of Kansas Medical Center

**Lemuel Russell Waitman[1], Gerald Lushington[2], Judith J. Warren[3]**

Biomedical Informatics accelerates scientific discovery and improves patient care by converting data into actionable information. Pharmacologists and biologists receive improved molecular signatures; translational scientists use tools to determine potential study cohorts; providers view therapeutic risk models individualized to their patient; and policy makers can understand the populations they serve. Informatics methods also lower communication barriers by standardizing terminology describing observations, integrating decision support into clinical systems, and connecting patients with providers through telemedicine.

The University of Kansas Medical Center's (KUMC) decision to pursue a National Institutes of Health (NIH) Clinical and Translational Science Award (CTSA)[1] in partnership with other regional academic medical centers, universities, and health systems catalyzed the development and integration of informatics capabilities to specifically support translational research in our region and to address issues and needs like those identified above. Headquartered at KUMC, the NIH-CTSA-supported Frontiers program, also called The Heartland Institute for Clinical and Translational Research (HICTR), is a network of scientists from institutions in Kansas and the Kansas City region (www.FrontiersResearch.org). The vision for the informatics section is to provide rich information resources, services, and communication technologies across the spectrum of translational research.

Broadly the initiative would: a) adopt methods which facilitate collaboration and communication both locally and nationally, b) convert clinical systems into information collection systems for translational research, c) provide innovative and robust informatics drug and biomarker discovery techniques, d) work with state and regional agencies to provide infrastructure and data management for translational outcomes research in underserved populations, and e) measure clinical information systems' ability to incorporate translational research findings.

The specific aims for informatics target translational needs requiring further investment and complement the novel methods and technologies of our region: a) the Institute for Advancing Medical Innovation's (IAMI)[2] effort in drug discovery and b) the Personalized Medicine Outcomes Center at the Uni-

*1 Director Medical Informatics, Department of Biostatistics, University of Kansas Medical Center*
*2 Associate Scientist and Director of Laboratories, Molecular Structures Group, University of Kansas*
*3 Christine A. Hartley Centennial Professor, School of Nursing, University of Kansas Medical Center*

versity of Missouri- Kansas City (UMKC) which seeks to understand the determinants of health systems' outcomes and develop methods to deliver evidence based practice. Four aims were articulated:

1. Provide a portal for investigators to access clinical and translational research resources, track usage and outcomes, and provide informatics consultation services.

2. Create a platform, HERON (Healthcare Enterprise Repository for Ontological Narration), to integrate clinical and biomedical data for translational research.

3. Advance medical innovation by linking biological tissues to clinical phenotype and the pharmacokinetic and pharmacodynamic data generated by research in phase I and II clinical trials (addressing T1 translational research).

4. Leverage an active, engaged statewide telemedicine and Health Information Exchange (HIE) to enable community based translational research (addressing T2 translational research).

This article will focus on Frontiers' plan and progress in achieving these aims since the grant was awarded in June 2011.

**Background**

The University of Kansas has a wide range of informatics capabilities, developed over decades of research and service. Through the National Center for Research Resources (NCRR) IDeA Networks for Biomedical Research Excellence (INBRE), the Kansas' KINBRE has created a ten campus network of collaborative scientists using common bioinformatics resources. Kansas has pioneered the use of telemedicine since the early 1990s, providing inter-state connectivity at over 100 sites and conducting thousands of clinical consultations and hundreds of educational events for health care professionals, researchers, and educators.

The Center for Health Informatics' Simulated E-hEalth Delivery System, a jointly funded program between the University of Kansas and Cerner Corporation, creates an equitable partnership to fully integrate applied clinical informatics into an academic setting and supports over 40 international academic clients. These existing efforts are complemented by more recent investments in clinical research informatics and medical informatics.

Starting in 2005, KU selected Velos e-Research for our Clinical Research Information System (CRIS). CRIS is a web application, supports Health Level 7 messaging for systems integration, and complies with industry and federal standards (CFR Part 11) for FDA regulated drug trials. It consists of a management component for patients, case report forms, and protocols, and allows financial management for conducting studies with administrative dashboards and milestones for measuring research effectiveness. KU's support and staffing serves not only support KU investigators but also to provide enterprise licensing for Frontiers' investigators to conduct multi-center, cooperative group, and investigator-initiated research with CRIS across unlimited participating locations without additional license fees. CRIS has been deployed throughout our institution and the team has experience sup-

porting multicenter trials across the nation.

Bioinformatics is a critical technology for translational (T1) research that systematically extracts relevant information from sophisticated molecular interrogation technologies. Analytical techniques–such as microarray experiments, proteomic analysis and high throughput chemical biology screening–can probe disease etiology, aid in development of accurate diagnostic and prognostic measures, and serve as a basis for discovering and validating novel therapeutics. HICTR institutions benefit from strong molecular bioinformatics research expertise distributed across three synergistic entities: the Center for Bioinformatics at KU-Lawrence (CBi), the Bioinformatics and Computational Life Sciences Research Laboratory (BCLSL), and the K-INBRE (Kansas IDeA Network for Biomedical Research Excellence) Bioinformatics Core (KBC). Over time, CBi (an internationally recognized structural biology modeling research program, focusing on computational structural biology) and BCLSL (a research program, led by the School of Engineering at KU-Lawrence and focusing on development of sophisticated algorithms for mining biological data) should grow into collaborative partners for Frontiers investigators.

The University of Kansas Center for Health Informatics (KU-CHI), established in 2003, is an interdisciplinary center of excellence designed to advance health informatics through knowledge integration, research, and education of faculty and students in the expanding field of biomedical science and information technology. Like many funded CTSA programs, KU is a founding member of the American Medical Informatics Association's (AMIA) Academic Forum. Graduate health informatics education, continuing education, AMIA 10x10 program, consultation, and staff development workshops/seminars designed to advance health informatics are sponsored or co-sponsored through KU-CHI. Graduate education in health informatics began in 2003 in the School of Nursing and is now a multidisciplinary master's degree program offered by the Office of Graduate Studies and managed by the KU-CHI. Helen Connors, PhD, RN, FAAN (executive director of KU-CHI) chairs the State of Kansas E-Health Advisory Council that oversees the development of the state's health information exchange strategic and operational plans and is a board member of the Kansas City Bi-state Health Information Exchange, the metropolitan area's Regional Health Information Organization.

Under the direction of Dr. Ryan Spaulding, The University of Kansas Center for Telemedicine & Telehealth (KUCTT) is a leader in telehealth services and research. The program began in 1991 with a single connection to a community in western Kansas. The Kansas telehealth network now is used to connect 60 health facilities in Kansas each year. It also is the basis for several national and international collaborations, demonstrating significant potential of telehealth technologies to eliminate distance barriers. Over the last 19 years, nearly 30,000 clinical consultations have been conducted across nu-

merous allied health, nursing and medical specialties, making the KUCTT one of the earliest and most successful telehealth programs in the world. In 2009, over 5000 Kansas patients benefited from telemedicine services that included clinical consultations, community education, health screenings and continuing education for health professionals. Through the Midwest Cancer Alliance (MCA) alone, the KUCTT supported numerous second opinion cancer consultations, multidisciplinary tumor board conferences, chemotherapy administration courses and routine patient consultations. Other specialties facilitated by the KUCTT include cardiology, psychology, psychiatry, neurology, wound care, etc. The KUCTT also has been very active integrating numerous technologies to support clinical and research activities, including telehealth systems, electronic health records, data registries and patient management systems.

In 2010, KU committed $3.5 million dollars over the next five years towards medical informatics academics and service, and established the Division of Medical Informatics in the Department of Biostatistics, to integrate clinical research informatics, and to provide overall leadership for the Frontiers Biomedical Informatics program. Coincidental with establishing the division, the University of Kansas Hospital completed their five-year, $50 million investment to implement the Epic Electronic Medical Record, concluding with Computerized Provider Order Entry in November 2010. This effort has transformed the inpatient environment. Subsequently, the University of Kansas Physicians began adoption of Epic across all ambulatory clinics.

**Planned Program Objectives and Progress.**

**Aim #1: Provide a Frontiers portal for investigators to access clinical and translational research resources and provide informatics consultative services.**

As shown at the top of Figure 1, a web application to link research resources and foster communication was seen as an underlying need. It provides investigators access to request services from the Translational Technologies and Resource Center, the Participant & Clinical Interactions Resources, the Community Partnership for Health, and the Ethics section. It also allows investigators to obtain targeted internal and extramural funding opportunities, request biostatistics support and clinical research database construction, request and access informatics resources, and explore educational offerings.

A key component of the portal is the adoption of REDCap[4] (Research Electronic Data Capture), to create databases to manage requests and triage review and tracking of requests. Medical Informatics implemented REDCap in January 2011 to provide a self-service database management system for clinical and translational research. Since users become familiar with the ease of use and development of forms, it was a natural approach to also extend the use of REDCap to manage research request activities. Thus, the user focused portal is complemented by using REDCap to help Frontiers management track projects, distribute resources, and facilitate evalu-

ation similar to efforts at Vanderbilt University[3]. Frontiers adoption of REDCap has been dramatic with over 700 users of over 500 projects either under development or in production. REDCap usage has ranged from student research projects as part of a fourth year Health of the Public medical school course to providing the data management infrastructure for a $6.3 million NIH funded Alzheimer's Disease Core Center grant.

In addition to the existing capabilities of the CRIS, there are several informatics specific components have been developed to provide better integration for clinical and translational researchers. Integrated authentication between the portal, the campus Central Authentication Service, CRIS, and i2b2 streamlines access to our clinical data repository described in the second aim. REDCap was used to develop secure methods for the process for submitting data requests and receiving data and complementary methods to support Data Request Oversight Committee activities. Finally, we provided a request management informatics consult service and training described below.

With the award of the CTSA, we anticipated the need to dedicate a clinical informaticist to match investigators' needs against current information available in Frontiers resources, specifically the HERON repository described in Aim 2. By early 2012, we had recruited Tamara McMahon to serve in this role as the clinical informatics coordinator. She assists users with tools like i2b2 and will recommend alternative approaches when required data are not integrated. She and the larger team's experience from front-line consulting with translational researchers drives prioritizing ad-

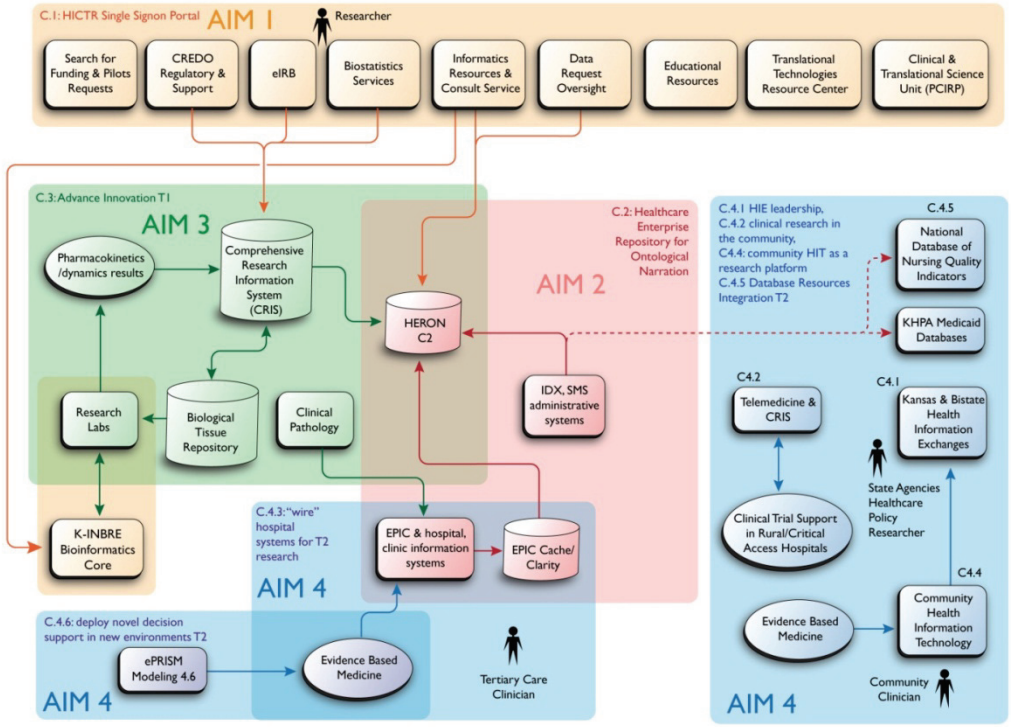



Figure 1. Conceptual Model of Biomedical Informatics and Specific Aims

ditional data sources, terminology and ontology development and guides the adoption of new technologies for data retrieval and analysis. As we've matured as a team these consultative services, while lead by the coordinator, are staffed and addressed by the combined biomedical informatics personnel. Additionally, informatics clinics have been established biweekly for investigators, staff, and students to attend and discuss needs and solutions in an informal group setting. Requests for bioinformatics and computational biology capabilities are facilitated by the Bioinformatics Resources section of the portal and support by the bioinformatics assistant director (initially Dr. Gerald Lushington; currently Dr. Paul Terranova).

**Aim #2: Create a platform, HERON (Healthcare Enterprise Repository for Ontological Narration), to integrate clinical and biomedical data for translational research.**

Transforming observational data into information and knowledge is the cornerstone of informatics and research. The opportunity for data driven knowledge discovery has exploded as health care organizations embrace clinical information systems. However, when dealing with confidential patient information, science and technology are dependent on law, regulation, organizational relationships, and trust. HERON provides Frontiers with secure methods for incorporating clinical data into standardized information aligned with national research objectives. It facilitates hypothesis generation, allows assessing clinical trial enrollment potential, and minimizes duplicate data entry in clinical trial data capture, outcomes research, and evaluation of T2 interventions for translating research into practice. While initially focused on integrating data between the KU Medical Center, the KU Hospital, and KU affiliated clinics, our methods are designed to subsequently integrate information among all Frontiers network institutions.

Beginning in April 2009, the Frontiers started piloting a participant registry in the clinics which allows patients to quickly consent to be contacted by clinical researchers. The registry contains the consented patients' demographics, contact information, diagnoses, and procedure codes. From April to July 2010, the medical center, hospital and clinics reviewed comparable practices[5,6,7,8] and drafted a master data sharing agreement between the three organizations that was signed September 6, 2010. The HERON executive committee is composed of senior leadership (e.g. chief operating, financial, executive officers and chief of staff) from the hospital, clinic and medical center and provides governance for institutional data sharing. Establishing business processes and servicing research requests is conducted by the Data Request Oversight Committee (DROC) which reports to the HERON executive committee. The repository's construction, oversight process, system access agreement, and data use agreement for investigators were approved by the Institutional Review Board. Since HERON is currently funded by KU and contains medical center and affiliated clinics and hospital data, access requires a medical center investigator. As additional institutions and health care organizations pro-

vide support and contribute data, they will be incorporated with multi-institutional oversight provided by the HERON executive committee and DROC.

As agreed upon by the HERON executive committee, HERON has four uses: cohort identification, de-identified data, identified data, and participant contact information:

After signing a system access agreement, cohort identification queries and view-only access is allowed and activity logs are audited by the DROC.

Requests for de-identified patient data, while not human subjects research, are reviewed by the DROC.

Identified data requests require approval by the Institutional Review Board prior to DROC review. After both approvals, medical informatics staff will generate the data set for the investigator.

Investigators who request contact information for cohorts from the HICTR Participant Registry have their study request and contact letters reviewed for overlap with other requests and adherence to policies of the *Participant and Clinical Interactions Resources Program* Data Request Committee.

The utility of data for clinical research is proportional to the amount of data and the degree to which it is integrated with additional data elements and sources. However, as additional data are integrated and a richer picture of the patient is provided, privacy concerns increase. HERON receives and preserves data into an identified repository, links and transforms those data into de-identified, standardized concepts, and allows users to retrieve data from this separate de-identified repository (See Figure 2). We recognize that certain research requires access to identified data but our approach streamlines oversight when needs are met with de-identified data. We adopted the NIH funded i2b2[9] software for user access, project management, and hypothesis exploration and chose Oracle as our database because of existing site licensing and i2b2 compatibility. Transformation and load activities are written in the python language and deployed onto SUSE LINUX servers. Servers are maintained in the medical center and hospital's data center which complies with HIPAA standards for administrative, technical and physical safeguards.

Maximizing data sharing to adhere to NIH guidelines has been a key goal for the development of HERON. Our goals are to make scientific data and informatics methods available as widely as possible while safeguarding the privacy of our patients. Our data sharing agreements were developed at an organizational level to streamline requests for de-identified data. Our repository's use of i2b2 and UMLS terminologies facilitate national collaboration. By using open source development tools and languages all new software developed using support from the NIH grant is then made available to others through our division's research and development website: http://informatics.kumc.edu.

Access to the identified repository is highly restricted and monitored. Data exchange between source systems, HERON, and user access is handled via security socket layer communications, https with digital certificates, and 128-bit

or higher public key cryptography (ex: SSH-2). We will continue to monitor federal guidance regarding appropriate security measures for healthcare information[10]. User access, activity logs, and project management for the de-identified repository is managed via the i2b2 framework and is integrated with the medical centers' open source Jasig Central Authentication Service and the HICTR portal. Account creation and access control is provided by the medical center. Audit logs are maintained for both the identified and de-identified repositories. Intrusion detection mechanisms are used and monitored by medical center Information Resources security personnel who also conduct HIPAA Security Rule reviews of the systems.

Current literature highlights challenges with providing anonymity after de-identifying data[11,12,13,14]. While we remove all 18 identifiers to comply with HIPAA Safe Harbor, the de-identified repository is treated as limited data set, reinforced by system access and data use agreements. Initially, HERON will only incorporate structured data. We will stay abreast of best practices (e.g., De-ID, the MITRE Identification Scrubber Toolkit versus only extracting concepts) prior to incorporating narrative text results which have greater risk for re-identification[11,15].

Arvinder Choudhary, our medical informatics project director, and Daniel Connolly, our lead biomedical informatics software engineer, were responsible for the initial construction of HERON. Starting in April 2010, the team has deployed a data repository with i2b2 integration, contributed to the i2b2 community code for integration with the Central Authentication Service (an open source project used by many universities for "single sign-on"), worked with other CTSA institutions to use the UMLS as a source terminology for i2b2, and established a software development environment for the team using Trac and Mercurial. Initial pilot use began as early as April for access to the Frontiers Participant Registry (diagnoses, demographics, and procedures based on clinic billing systems). The current architecture was deployed in August 2010 and populated with production data from IDX and Epic in September 2010.

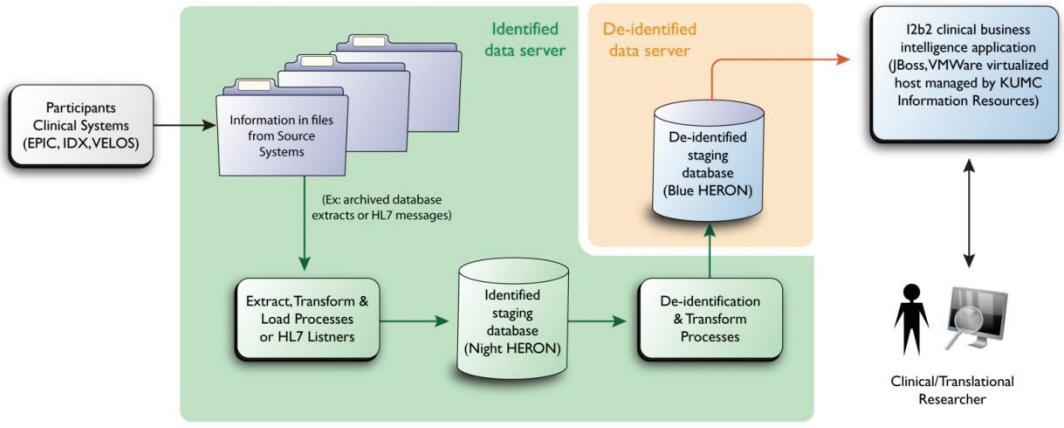



Figure 2. HERON Architecture

Initial extract, load, and transform processes obtain data files from the source system via secure file transfer methods. The repository was available as a proof of concept in December 2010 and entered beta mode for all faculty in March 2011. HERON has been updated with new data and functionality every month since August 2011 (updates advertised on our blog: http://informatics.kumc.edu/work/blog). As of August 2012, HERON is using an i2b2 version 1.6 and it contains approximately 850 million facts for 1.9 million patients.

From its inception, HERON has benefitted from i2b2 community expertise and the team works with other academic medical centers to align our terminology with CTSA institutions. Our approach is to rely on the National Library of Medicine Unified Medical Language System (UMLS) as the source for ontologies and to use existing or develop new update processes for the i2b2 Ontology Management Cell and other systems. The National Center for Biological Ontologies is also working with the i2b2 development team to create methods that can assist organizations using i2b2 with building appropriate ontologies based on the UMLS or other source ontologies. This may play a greater role as we bridge clinical data towards bioinformatics domains and to incorporate more recently authored ontologies that are not fully supported by the NLM. As we develop new mapping for going from UMLS into i2b2 and from source systems such as Epic, we have shared

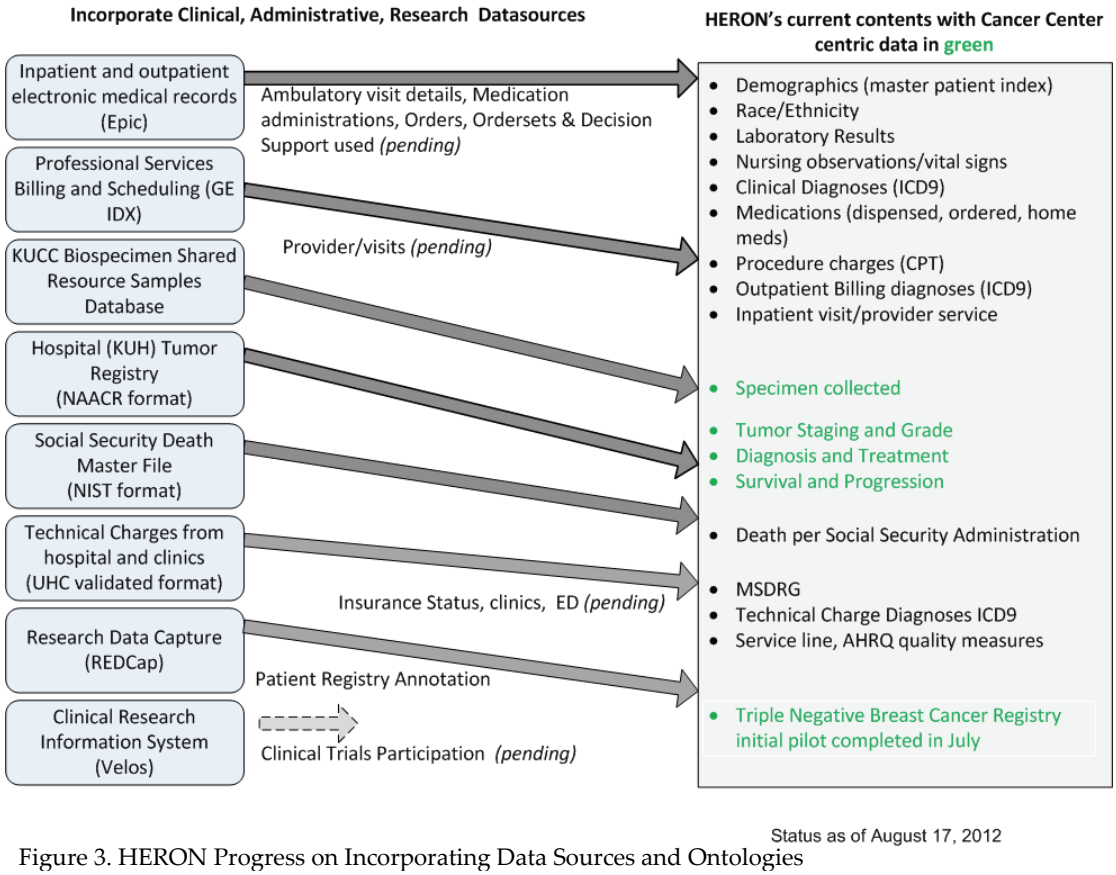



Figure 3. HERON Progress on Incorporating Data Sources and Ontologies

our code and processes with the i2b2 community via our website. Figure 3 illustrates our current progress in building a rich picture of our patient population and highlights places where the HERON repository also serves to advance KUMC's goal to become a National Cancer Institute designated Cancer Center.

The team's approach leveraged methods outlined by several CTSA organizations that shift mapping inside i2b2 instead of in database transformation and load processes[16]. This preserves the original data side by side with the transformed information allowing easier review by investigators and domain experts. We applied these methods when we incorporated nursing flowsheet information[17] (vital signs, nursing assessments, social status) from the Epic electronic medical record. Judith Warren, PhD, RN, BC, FAAN, FACMI (Director of Nursing Informatics, the Center for Health Informatics) serves as assistant director of health informatics for Frontiers and has extensive experience developing health informatics standards and terminologies, teaching informatics, and developing national policy for health information technology. She has led several initiatives, especially regarding nursing observations, to adopt standards that will maximize data integration and reuse.

Our hope is that standardizing information for research strengthens our relationship with the hospital and clinics. Reusing information for clinical quality improvement also requires

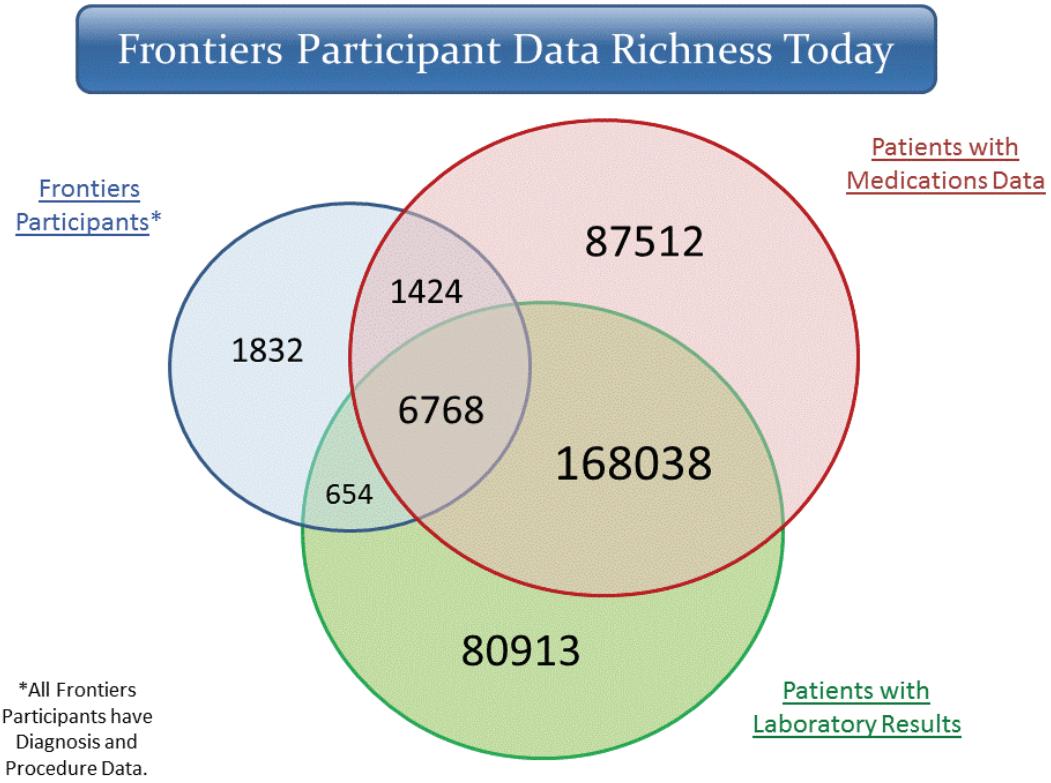



Figure 4. Richness of Phenotype in HERON using Frontier Participants as an example

standardization. Translational research using HERON has highlighted areas needing terminological clarity and consistency in the electronic medical record. Going forward, we are partnering with Dr. Ator, Chief Medical Information Officer, and the University of Kansas Hospital's Organizational Improvement (which has two of their leaders on the DROC) to develop a data driven process for improving standardization. Research and quality improvement goals also align with achieving "meaningful use"[18] and Health Information Exchange activities. We will gradually build relationships by initially focusing on linking data sources from KU Hospital and affiliated clinics. Figure 4 provides an illustration of how today, patients who volunteer to be contacts for research in the clinic based upon a clinic billing system indicator are linked to medication exposure and laboratory results, allowing more targeted recruitment for prospective clinical trials[19]. The majority of these patients have diagnoses, demographics, procedure codes, laboratory results, and medications in HERON. As capabilities of the team develop we plan to expand to other Frontiers affiliated network institutions, state agencies, and our rural practice networks (outlined in aim 4). This may require centralized and/or distributed data integration[20,21].

**Aim #3: Advance medical innovation by linking biological tissues to clinical phenotype and the pharmacokinetic and pharmacodynamic data generated by research in phase I and II clinical trials (addressing T1 translational research).**

We saw two points where Frontiers' informatics must bridge between basic science resources and clinical activities to enhance our region's ability to conduct translational research. Our region provides a wealth of biological and analytical capabilities but struggles at times to understand if our clinical environments see enough patients to support research. This leads to clinical trials which fail to accrue enough subjects[22]. Accurately annotating biological tissues and routine clinical pathology specimens with the patients' phenotypic characteristics (such as diagnosis and medications from the clinical records) would provide a more accurately characterized clinical research capacity. Our second targeted area is designed to support the Drug Discovery and Development activities of Frontiers' Institute for Advancing Medical Innovation's (IAMI) initiative in managing information in phase I and II clinical trials. Molecular biomarker and pharmacokinetics/pharmacodynamics research activities provide high quality analysis, but their results are not easily integrated with other records maintained in clinical research information systems. Reciprocally, if clinical characteristics could be provided with samples in a standardized manner it would allow improved modeling and analysis by the core laboratories. Collaboration between the bioinformatics and biomedical informatics specifically targeted at biospecimens, pharmacokinetic /pharmacodynamic results and clinical trials will provide a foundation for subsequent integration and our final goal of providing molecular bioinformatics methods.

Aligning CTSA capabilities with our Cancer Center's pursuit of national designation was a top T1 priority for the informatics team. Our ability to incorporate clinical pathology based information and the biological tissue repositories with HERON could improve cohort identification, clinical trial accrual, and clinical trial characterization. Patients' tissue specimens are a fundamental resource that links the majority of basic biological science analysis to clinical relevance. We need to improve our characterization of these samples to maximize our investment in maintaining them for clinical research. By integrating clinical pathology derived characteristics and research specimens with clinical information in HERON, researchers can quantify samples belonging to patients (e.g. aggregating all breast cancer diagnoses using the i2b2 clinical concept "malignant neoplasm of the breast" and treated with the adjuvant therapy Trastuzumab).

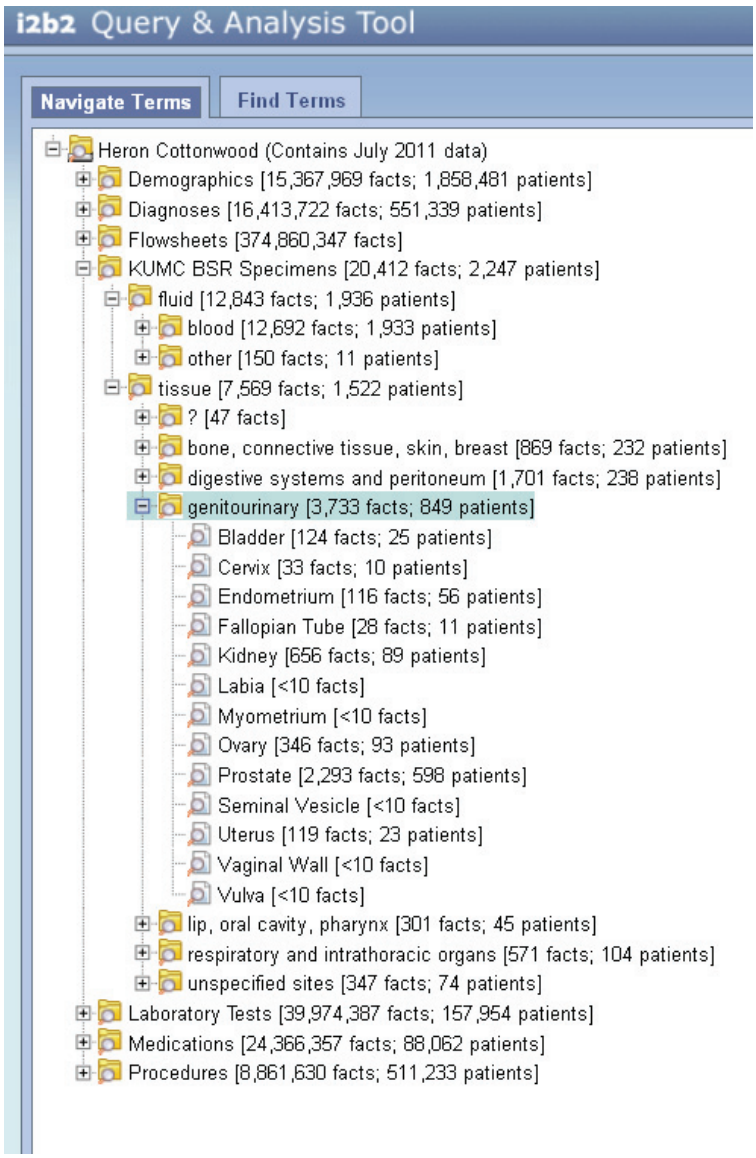



Figure 5. Initial ontology for Biospecimen Tissue Repository Integration within HERON

On July 12, 2012, KU was awarded NCI Cancer Center Designation, the highest priority strategic objective for the university. Informatics contributed by enhancing i2b2 for cancer research by incorporating key data sources and developing methods for conducting real time survival analysis. Our first step towards supporting T1 research was integrating the Biological Tissue Repository (BTR) in the *Translational Technologies and Resource Center* (Section 8) and HERON. The BTR, built upon the existing KU Cancer Center Biospecimen Shared Resource, plays a vital role in collecting and distributing high quality human biospecimens essential to research. Improving specimen management for clinical research has been identified as a high

priority initiative for KU. Integrating the BTR allows users to exploit the phenotypic data in HERON to characterize the samples. BTR personnel can also use this linkage to provide investigators requesting samples with clinical phenotype from HERON via a data use request approved by the oversight committee. Figure 5 demonstrates the ontology for biospecimens within HERON's i2b2 application. We are currently migrating the BTR to reside within the Comprehensive Research Information System (CRIS). This will enhance analysis and characterization of data from clinical trials and also improve inventory management, retrieval, and quality assurance processes for the BTR staff.

Our second step was to incorporate the hospital's Tumor Registry within HERON since the validated tumor registry provides additional tumor characteristics for the cancer population, outcomes, and follow up information that are not maintained in surgical pathology systems. The tumor registry was incorporated with collaboration from Kimmel Cancer Center (PA) and Group Health Cooperative (WA) and leveraged using the national standard for tumor registry data specified by the North American Association for Central Cancer Registries (NAACCR). We also incorporated the social security death index from the National Institutes for Standards and Technology to provide support for long term outcomes and survival analysis across all clinical conditions and health services research. Finally, we created rgate – an improved method to integrate R with i2b2 and Oracle[23]. Figure 6 provides a summary of HERON workflow for hypotheses generation and preliminary visualization using the interactive analysis tools. Future work will focus on generalizing plugins for multiple cohorts beyond cancer and extending the use of R analysis within HERON so that initial validation can be conducted without needing to request data from HERON.

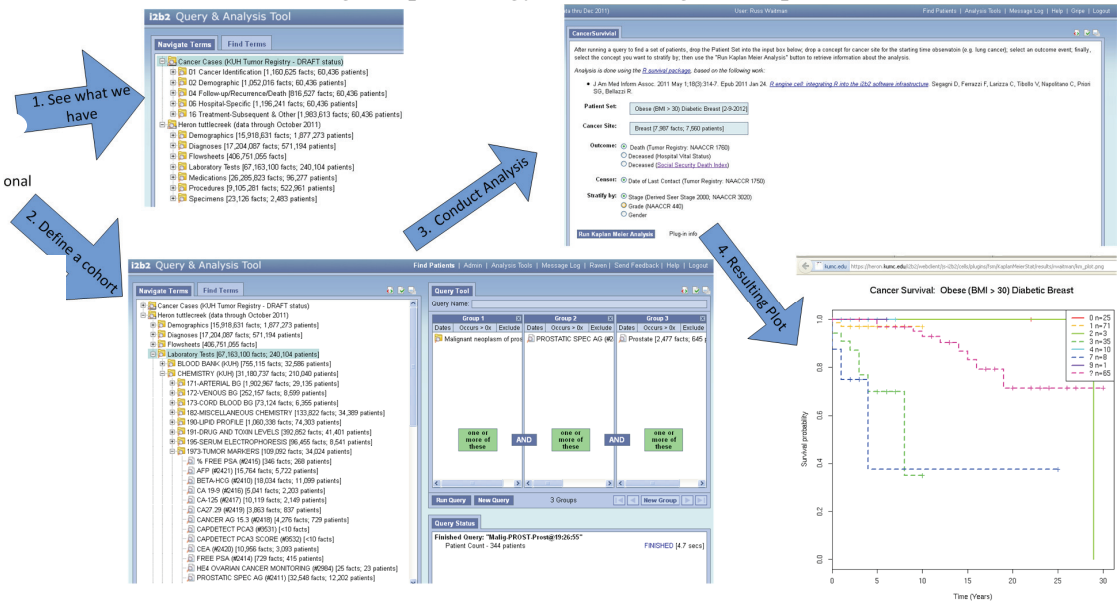



Figure 6. HERON providing interactive survival analysis

That in turn reduces the risk of accidental data disclosure.

Going forward, we plan to support IAMI clinical trials by integrating and standardizing information between the bioinformatics and pharmacokinetic/pharmacodynamic (PK/PD) program and CRIS. The Institute for Advancing Medical Innovation (IAMI) builds upon the region's significant drug discovery capabilities and creates a novel platform for medical innovations. Members of the K-INBRE Bioinformatics Core (KBC) provide bioinformatics analysis across the spectrum of target discovery, compound screening, lead identification and optimization, and preclinical development. As drug candidates successfully transition to a new investigational drug application, we will continue bioinformatics support into clinical trials.

Through Frontiers we hope to establish stronger bridges between analytical resources and clinical information which will allow phenotypic variables to be exploited by bioinformatics expertise in data mining by the K-INBRE Bioinformatics Core (KBC) and the Bioinformatics and Computational Life Sciences Research Laboratory. Over time, Frontiers T1 translational research efforts may be enhanced by providing data mining algorithms that extract key information from data-intensive wet lab studies and integrating those results with phenotype derived from clinical applications. This bridge between molecular-level analysis and clinical studies is a potent model for biomarker prediction and validation, PK/PD studies, and the prospective refinement of personalized medicine.

**Aim #4: Leverage an active, engaged statewide telemedicine and Health Information Exchange (HIE) to enable community based translational research (addressing T2 translational research).**

In the latter years of the grant we plan to leverage regional informatics capabilities to complement the long history of engaging the community through outreach via telemedicine, continuing medical education, and our extensive community research projects in urban rural and frontier settings. We will highlight the initial areas where Frontiers is deploying informatics in support of T2 community based research.

Under the leadership of Dr. Helen Connors, the Center for Health Informatics has been instrumental for advancing HIE for Kansas and the Kansas City bi-state region. As chair of the eHealth Advisory Council, Dr. Connors guided the process to establish the Kansas HIE (KHIE)[23]; a non-profit corporation incorporated by the governor in response to the federal HITECH Act. Allen Greiner, MD, director of the *Community Partnership for Health* (Section 5), participates in selecting electronic health records for the Kansas Regional Extension Center and Dr. Russ Waitman has participated in activities on behalf of Kansas Medicaid and technical subcommittees responsible for the development of Health Information Exchange in the state. The shared goal is to keep Frontiers engaged in state and regional infrastructure capabilities for translational research and facilitate identifying other collaborating organizations.

The University of Kansas Center for Telemedicine and TeleHealth (KUCTT) provides enormous reach into diverse communities for implementing clinical trials and other research across the state. Telemedicine has figured prominently in a number of clinical research programs (e.g. studies of diagnostic accuracy, cost effectiveness, and patient and provider satisfaction with telemedicine) as well as studies that employ telehealth for randomized control trial interventions. Complementing KUCTT, the Midwest Cancer Alliance (MCA) is a membership-based organization bringing cancer research, care and support professionals together to advance the quality and reach of cancer prevention, early detection, treatment, and survivorship in the heartland. The MCA links the University of Kansas Cancer Center with the hospitals, physicians, nurses and patients battling cancer throughout Kansas and Western Missouri. The MCA advances access to cutting-edge clinical trials as well as professional education, networking, and outreach opportunities. These clinical trials use our Comprehensive Research Information System (CRIS). Notably, the MCA critical access and rural referral centers are also the most active participants in telemedicine. The medical director of the MCA, Dr. Gary Doolittle is a pioneer user and has published telemedicine research with Dr. Ryan Spaulding, the director of the telemedicine network[24,25]. The network has been successfully used to facilitate clinical research and to support special needs populations; especially in the area of smoking cessation.

Biomedical informatics works with KUCTT to provide a telemedicine clinical trials network connected by CRIS for remote data collection via study personnel or direct data entry through a patient portal.

Additionally, in March 2011, Kansas City, Kansas "won" a nationwide Request for Information (RFI) solicitation for construction of a next generation broadband network by Google achieving upload and download speeds of 1 gigabit per second (Google 2011). Google subsequently announced the expansion of service to Kansas City, Missouri for a potential deployment of 500,000 individuals. Through Frontiers, the University of Kansas Medical Center's (KUMC) early support led to a collaborative working relationship between KUMC Medical Informatics and the Google Fiber team to explore leveraging the technology to further health care delivery, education, and research. On July 26, 2012, Google unveiled their services and began preregistration throughout the Kansas City metropolitan area. KUMC is actively participating by providing nutritional consults via telepresence as part of the Google Fiber Space located a few blocks from the medical center campus. Building on the strength of telemedicine and informatics Drs. Waitman and Spaulding are collaborating with the School of Nursing and the School of Engineering to develop a research proposal for the National Science Foundation's US Ignite (www.nsf.gov/usignite) initiative to develop next generation applications that can exploit ultra-fast networks such as Google Fiber.

In the coming year, Frontiers plans to build upon investment in electronic medical records by consulting with Frontiers investigators to optimize the use of clinical systems for disseminating translational evidence and recording measures to verify adoption. This work is also foundational to medical informatics and will allow the evaluation of native clinical information system "signals" against manual processes used to report measures to government, regulatory agencies, and national registries. Working with the hospital to develop capabilities to acquire user activity data to measure the systems' impact on workflow and clinical decision making (e.g. drug-drug interaction overrides, time to complete medication administration task) will also help us overcome translational research's last mile: implementation into clinical workflow.

Longer term objectives for Frontiers will look to develop methods to engage with community providers regarding electronic health record adoption and those systems' capacity to support translational research. As mentioned previously, we have also made initial inroads regarding incorporating state and national data sources into HERON as illustrated by our incorporation of the Social Security Death Index. Over time, we hope to work with the Kansas Medicaid program to investigate methods to link health data maintained by the state as claims against the clinical data in HERON. This will further three objectives: (a) national data may provide outcome measures lacking in acute care clinical information systems, (b) national registries, such as the National Database of Nursing Quality Indicators, can evaluate the degree to which manually abstracted measures might be automatically derived from a mature clinical information system, and (c) data integration will allow hypotheses to be explored that suggest methods for improving care. Finally, we aim to collaborate with Frontier's Personalized Medicine and Outcomes Center to provide complex risk models for decision support to a variety of clinical specialties. Those investigators have developed methods that translate complex risk models into fully functional decision-support tools for physicians as well as personalized educational and informed consent documents for patients.

**Acknowledgements**

Large portions of this text were previously used in the university CTSA application's biomedical informatics section by the coauthors Lemuel Waitman, Judith Warren and Gerald Lushington. We would like to acknowledge the participation of Frontiers leadership and numerous investigators throughout the Kansas City region who have contributed to the development and ongoing progress of Frontiers.

**References**

1. Zerhouni EA. Translational and clinical science–time for a new vision. NEJM 2005, 353:1621-23.
2. The University of Kansas Medical Center 2012. Institute for Advancing Medical Innovation. Retrieved from http://www.kumc.edu/iami.html
3. Pulley JM, Harris PA, Yarbrough T, Swafford J, Edwards T, Bernard GR. Acad Med. 2010 Jan;85(1):164-8. An informatics-based tool to assist researchers in initiating research at an academic medical center: Vanderbilt Customized Action Plan.
4. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap) - A metadata-driven methodology and workflow process for providing translational research informatics support, J Biomed Inform. 2009 Apr;42(2):377-81.
5. Sears CS, Prescott VM, McDonald CJ. The Indiana Network for Patient Care: A Case Study of a Successful Healthcare Data Sharing Agreement.

American Bar Association Health eSource, September 2005.
6. Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, Churchill S, Kohane IS. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc. 2009 Sep-Oct;16(5):624-30.
7. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, Masys DR. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther. 2008 Sep;84(3):362-9. Epub 2008 May 21.
8. Pulley J, Clayton E, Bernard GR, Roden DM, Masys DR. Principles of human subjects protections applied in an opt-out, de-identified biobank. Clin Transl Sci. 2010 Feb;3(1):42-8.
9. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010 Mar-Apr;17(2):124-30.
10. SP 800-66 Rev 1 Oct 2008 An Introductory Resource Guide for Implementing the Health Insurance Portability and Accountability Act (HIPAA) Security Rule. http://csrc.nist.gov/publications/nistpubs/800-66-Rev1/SP-800-66-Revision1.pdf
11. Friedlin FJ, McDonald CJ. A software tool for removing patient identifying information from clinical documents. J Am Med Inform Assoc. 2008 Sep-Oct;15(5):601-10. Epub 2008 Jun 25.
12. Loukides G, Denny JC, Malin B. The disclosure of diagnosis codes can breach research participants' privacy. J Am Med Inform Assoc. 2010 May 1;17(3):322-7.
13. Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. J Am Med Inform Assoc. 2010 Mar 1;17(2):169-77.
14. Karp DR, Carlin S, Cook-Deegan R, Ford DE, Geller G, Glass DN, Greely H, Guthridge J, Kahn J, Kaslow R, Kraft C, Macqueen K, Malin B, Scheuerman RH, Sugarman J. Ethical and practical issues associated with aggregating databases. PLoS Med. 2008 Sep 23;5(9):e190.
15. Yeniterzi R, Aberdeen J, Bayer S, Wellner B, Hirschman L, Malin B. Effects of personal identifier resynthesis on clinical text de-identification. J Am Med Inform Assoc. 2010 Mar 1;17(2):159-68.
16. Wynden R, Weiner MG, Sim I, Gabriel D, Casale M, Carini S, Hastings S, Ervin D, Tu S, Gennari J, Anderson N, Mobed K, Lakshminarayanan P, Massary M, Cucina R. Ontology mapping and data discovery for the translational investigator. Presented at the 2010 AMIA Summit on Clinical Research Informatics meeting, March 12-13 2010, San Francisco, CA.
17. Waitman LR, Warren JJ, Manos EL, Connolly DW. Expressing observations from electronic medical record flowsheets in an i2b2 based clinical data repository to support research and quality improvement. AMIA Annu Symp Proc. 2011;2011:1454-63.
18. Blumenthal D, Tavenner M. The "Meaningful Use" Regulation for Electronic Health Records. N Engl J Med. 2010 Jul 13.
19. Denton, J., Blackwell K., Waitman, L., Kluding, P., Choudhary, A., Bott, M., Brooks, W., Greiner, A., Jamison, R., Klaus, S., O'Brien-Ladner, A., Latinis, K., Aaronson, L., Burns, J., & Barohn, R. (2012). P334 Frontiers Research Participant Registry. Translational Science 2012 Meeting, Research Professionals Abstracts, Clinical and Translational Science, 5:2, http://onlinelibrary.wiley.com/doi/10.1111/j.1752-8062.2012.00398.x/full
20. Diamond CC, Mostashari F, Shirky C. Collecting and sharing data for population health: a new paradigm. Health Aff (Millwood). 2009 Mar-Apr;28(2):454-66.
21. Berman JJ. Zero-check: a zero-knowledge protocol for reconciling patient identities across institutions. Arch Pathol Lab Med. 2004 Mar;128(3):344-6.
22. Nass SJ, Moses HL, Mendelsohn. A National Cancer Clinical Trials System for the 21st Century: Reinvigorating the NCI Cooperative Group Program. Committee on Cancer Clinical Trials and the NCI Cooperative Group Program; Institute of Medicine. Washington, DC: National Academies Press, 2010.
23. Connolly DW, Adagarla B, Keighley J, Waitman LR. Integrating R efficiently to allow secure, interactive analysis within a clinical data warehouse. The 8th International R User Conference. Vanderbilt University; Nashville, Tennessee, USA June 12-15, 2012.
24. Spaulding RJ, Russo T, Cook DJ, Doolittle GC. Diffusion theory and telemedicine adoption by Kansas health-care providers: critical factors in telemedicine adoption for improved patient access. J Telemed Telecare. 2005;11 Suppl 1:107-9.
25. Doolittle GC, Spaulding RJ. Defining the needs of a telemedicine service. J Telemed Telecare. 2006;12(6):276-84.

# Creating Infrastructure for Research Collaboration

**Arun K. Somani**, Anson Marston Distinguished Professor, Electrical and Computer Engineering, Iowa State University

In the world of today, information is key to success. Plenty of research done today requires and/or uses exa-scale computers. High performance computing (HPC) is essential for advanced research in virtually all disciplines of science and engineering, and in particular in the fields of bio-, materials-, and information-technologies, all identified as strategic priorities of Iowa State University. The remarkable increase in computational capability (more broadly cyber infrastructure) over the last 40 years has changed societies, disciplines, and governments. Availability of contemporary HPC services is instrumental in researchers' ability to improve research quality and competitiveness. They are also essential in attracting and retaining top faculty working in vital disciplines that rely on the computational sciences, build strong inter-institutional research collaborations, and lead to new discoveries by enabling these researchers to scale up models beyond the known edges of prior work. A considerable investment in new HPC is thus crucial.

With these thoughts in mind, how can Iowa State University best position itself to optimize the use and development of cutting-edge HPC and to lead the changes? A University HPC Committee (U-HPC-C) was formed to address the need of HPC users at ISU campus. Our model to address the said needs included multi-faceted aim to achieve economies of scale in space utilization, to maximize the use of physical facilities, to leverage available funds, and to move from an ad hoc to a planned approach to support most medium to large scale users. **We believed that a strong case can be made for the acquisition of a new HPC platform that would satisfy the needs of the projects described herein, and provide sufficient capacity to meet the needs of a broad number of important research groups on campus.** The merit of the new HPC platform includes the far-reaching and impactful research thrusts that it enables, by accelerating knowledge discovery, spanning areas as diverse as animal sciences, plant genomics, climate modeling, and wind power generation.

ISU's computer users can be separated into two communities: i) those that do science and HPC is an instrument for them, and ii) those that do science to advance the state-of-the-art in HPC by developing new algorithms, architectures, data storage techniques, and management of computation. These two communities and their HPC needs are superficially different, but deep down both communities benefit and have benefitted from the strong interaction between them. *A guiding principle of our goal is to ensure that the first community benefits*

*from the new instrument to the fullest possible extent while facilitating time for HPC innovation by the second community.* We also believed that the new HPC platform required 100-200 TFlops compute capability, large memory (up to 16GB per core) and 1,000+ TB storage, and strong support for parallel I/O. The new HPC platform would handle the storage needs of various applications.

**Science and Its Need**

***We identified the following major projects that would benefit from such a facility.***

**Project 1. Biosciences.** ISU has large, well established interdisciplinary research and education programs at the forefront of biological sciences. These research efforts span microbial, plant and animal species, and are fully integrated with teams of computational scientists due to the transformation of biology as a data-driven science. In the microbial arena, ISU researchers are engineering microbial organisms for biorenewable production of chemicals and important energy-related compounds such as hydrocarbons. ISU has a preeminent program in plant sciences research with particular emphasis on biotechnology of important cereal crops such as maize, barley and soybean. Our comprehensive plant sciences research spans food, feed, fiber and fuel through eight research centers, along with environmental research through the Center for Carbon Capturing Crops. ISU animal and veterinary science researchers are engaged in genomics and systems biology of livestock for effective breeding, improving quality and nutrition of food, and study of diseases that affect livestock.

Though these applications are broad and diverse, fundamental advances in genomics and the common genetic mechanisms underpinning all life forms provide many cross-cutting synergies among these seemingly disparate fields of research. In particular, data-intensive experimental equipment is commonly used, including sequencers, microarrays, and mass spectrometers to measure metabolic fluxes. Thus, the computational and bioinformatics tools needed to drive such research share common methodologies and computing needs. The emergence of high-throughput DNA sequencing technologies and their rapid proliferation and throughput gains during the past five years has created a compelling need for the HPC equipment.

Sequencing individual plants and animals[1] from previously sequenced species (i.e., a template genome exists) enables determining structural variation (indel, SNP, CNV, re-arrangements) along with annotation. The animal sciences group is planning tens to hundreds of individual sequences for the bull, cow, pig, and chicken genomes, roughly the same size as the ~3 billion bp human genome. As part of the ISU Beef Research herd, researchers are sequencing 100 plus bulls in the next year followed by 5-10 bulls/year thereafter. For a nominal 50X sequencing coverage through short reads and 100 individuals, the data size is approximately 15 TB. Similarly, the plant sciences group is engaged in the sequencing of multiple varieties of crop plants. In addition to

storage and computational needs, genomic rearrangements are common in plants due to selective breeding, which creates short evolutionary history.

**Genome assembly**. Genome assembly[2] is the problem of inferring the unknown genome of an organism from a high coverage sampling of short reads from it obtained from a high-throughput sequencer. Assembly is needed when sequencing a new species, or when sequencing a new individual of a known species where genomic modifications make the use of a template genome unviable. ISU researchers are working on engineering *E. coli* for production of hydrocarbons.

Genome wide association studies enable identification of Single Nucleotide Polymorphisms (SNPs), which are single base differences between genomes of individuals from the same species. One use would be to link groups of SNPs to a phenotype such as a particular disease. Modern genome wide association studies are conducted using SNP chips, which consist of hundreds of thousands of high density probes allowing many SNPs to be interrogated simultaneously. For example, the animal sciences group is currently in the process of migrating from 50K SNP chips to 500K SNP chips as they are becoming available for cattle, swine and chickens. It is envisioned that whole genome data (3 million plus markers per individual) will be available in the near future (see Genome Re-sequencing section above). The goal is to perform interaction studies using all of these genes, which requires fitting a Markov chain of typically 50,000 cycles. Computationally, this results in a linear system of 500,000 equations that need to be repeatedly fit 50,000 times in order to compute the relevant posterior distributions as per the Bayesian approach.

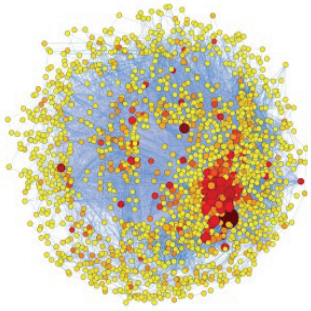


**Figure 1.** Backbone of Arabidopsis genome network containing 1,556 genes and 22,073 interactions. Display represents a union of the shortest paths between each pair of the top 5% of the hubs in the entire genome network. The size of a node is proportional to its degree and the intensity of its color is proportional to its betweenness centrality.
Picture Courtesy S. Aluru.

**Biological Network Inference and Analysis.** Biological networks[3] (see Figure 1) represent interactions between genes, RNAs, proteins and metabolites whose collective system-level behavior determines the biological function. Inference and analysis of networks depicting flow of activity and regulation are fundamental to understanding biological processes. Such knowledge is vital to the pursuit of engineering efforts to achieve a desirable outcome or to understand or treat disease (how does a malfunctioning gene alter the pathway?). ISU researchers have developed methods for gene network inference and analysis at the whole-genome scale, and protein interaction networks at the genome-scale. A typical data set for gene networks involves thousands of experiments, each providing expression values of every gene in the organism (~20,000 – 50,000 for plants), resulting in as many as 100 million values or more. The data

can be from microarray experiments or the newer RNA-seq experiments via next-gen sequencing. Inference is carried out via parallel methods that use sophisticated non-linear approaches.

**Metabolomics and *in Silico* Modeling.** Developing a comprehensive understanding of a biological system also requires the processing and integration of a large number of datasets from a variety of high throughput instrumentation, beyond just sequencers. ISU is home of several large databases funded by the NSF and USDA that analyze and make these datatypes available to the research community (e.g., PLEXdb, MaizeGDB, PlantGDB, PlantMetabolomics.org, and animalgenome.org). Several ISU researchers are engaged in metabolic flux analysis, a powerful diagnostic tool enabling quantification of all steady state intracellular fluxes in a metabolic network. The resulting maps provide a measure of the extent of contribution of various pathways in cellular metabolism. Another major research effort is to build genome-scale *in silico* models of organisms to be able to conduct flux analysis and ask "what if" questions to study the effect of genetic manipulations on desired end goals such as increasing the production of a metabolite.

**Project 2: Multiscale Methods for Grand Challenge Problems**

Methods that can accurately and efficiently span multiple length and time scales are required to address many "grand challenge" problems, such as design of new materials for specific applications, capture of solar energy, study of heterogeneous catalysis, simulation of the processes for biomass conversion to usable energy, simulation of atmospheric phenomena such as aerosol formation and reactivity, and analysis of enzyme catalysis. ISU researchers' goal is to develop methods that are capable of providing both accuracy and computational efficiency. Developing such methodology starts with high-level quantum mechanics (QM) methods that are computationally expensive and mapping the high-level potential onto a potential that is much simpler and much less computationally demanding, with minimal loss of accuracy. This process is called coarse graining. It divides a large molecular system into smaller, more computationally tractable fragments so that the properties of each fragment can be computed on a separate node, while the accuracy is not significantly compromised. This fragmentation scheme[4] greatly reduces the cost of a QM calculation and facilitates multi-level parallelism. The advantage of such a method and some remaining challenges can be illustrated by considering a large cluster of water molecules. Because water is arguably the most important liquid and solvent, to perform a realistic molecular dynamics (MD) simulation on water, one needs to start with ~1,024 waters molecules. Using 131,000 BG/P cores, a single energy + gradient calculation on 1,024 molecules requires 1.2 minutes. A realistic MD simulation with 1 million time steps, this time translates into 1,200,000 minutes = 2.28 years! One way to address the challenge presented by the steep scaling of QM methods is to map (coarse grain) the FMO potential onto a much simpler potential. Then one can employ graphical processing unit (GPU) accelerators.

**Formation and Reactions of Atmospheric Aerosols.** An understanding of the effect aerosols[5] have on the climate (and global climate warming) has become increasingly important over the last several decades. Primary and secondary aerosols affect Earth's radiative balance by scattering and absorbing light directly, and act indirectly as cloud droplets to influence the distribution, duration, precipitation processes, and radiative properties of clouds. It is also recognized that the spatial and temporal distributions of aerosols due to industrial activities are as important in determining overall climate changes as is the influence of greenhouse gases. Understanding the nucleation, growth and evaporation rates of aerosols as well as their chemical properties is essential to improve climate models and overall global climate prediction as well as the general chemical ecosystem in the atmosphere. Gaining this understanding is extremely challenging both from a scientific and a computational science point of view.

**Design of Dendrimers.** Dendrimers are a class of polymer with regularly branching repeat units emanating from a central core (Figure 2). They are synthesized using a series of controlled reaction steps, which endows these high molecular weight molecules with the structural precision of a small organic molecule. This attribute gives dendrimers an advantage over other polymers in biomedical applications, where strict regulatory requirements are imposed on polymer-based materials for use in humans. Applications for dendrimers include microbicides, drug and gene delivery, tissue engineering, imaging, and water and soil remediation. Designing dendrimers with specific materials properties requires a thorough understanding of how changes in the size, shape, and surface chemistry of a dendrimer affect its interactions with target species, such as low molecular weight organic drug molecules or industrial pollutants. Molecular simulation with atomistic resolution is an invaluable counterpart to experimental observations, provided that realistic and accurate molecular models are available.

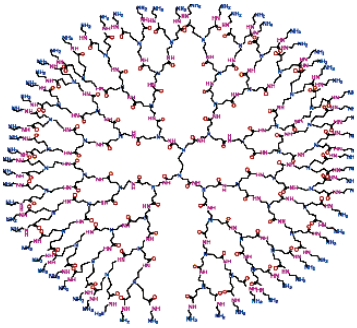


**Figure 2.** Structure of dendrimer showing core-shell structure.

**Project 3: Computational Fluid Dynamics (CFD) Modeling**

**Introduction.** The CFD modeling group at ISU makes extensive use of high-performance computing to carry out cutting-edge research using simulation methods in fluid mechanics and multiphase flows[6]. These simulations will fundamentally advance the century-old challenge: the direct computation of turbulent flows at flight conditions. Each simulation requires about 40 processors with a memory of 1GB per processor and the run time of a typical simulation is about 10 days. Particle-resolved DNS simulations of large risers in three dimensions have not been performed previously. Algorithmic developments made in designing optimal parallelization strat-

egies for particle-resolved DNS with fluid-particle coupling have broader applications in sprays, bubbly flows and device-scale simulations of gas-solid flow applications that employ discrete element methods to treat the solid phase.

**Fundamental Physics of Multiphase Reactive Flows.** Multiphase reactive flows are encountered in energy generation processes using fossil or bio-based fuels and in emerging technologies to capture the resulting $CO_2$ emissions. Technologies such as chemical-looping combustion and capture of $CO_2$ using dry sorbents promise to reduce greenhouse gas emissions. CFD plays an important role in the design and scale-up of these devices. The multiphase flow in these industrial devices is characterized by complex hydrodynamics, heat transfer and both exothermic and endothermic chemical reactions. A fundamental understanding of the interaction between hydrodynamics, heat transfer and chemistry over a wide range of physical parameters (such as solid phase volume fraction, Reynolds number) is crucial for successful device-scale CFD models that can be used to design systems for these emerging technologies. ISU researchers have developed a method that represents the contact mechanics of multiparticle collisions between particles up to the close-packing limit of solid volume fraction (Figure 3), requiring use of high-performance computing.

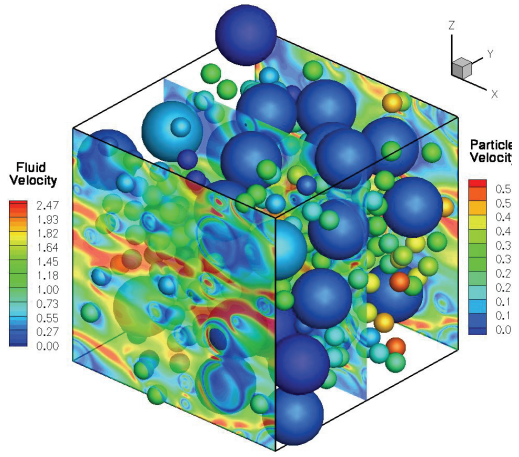
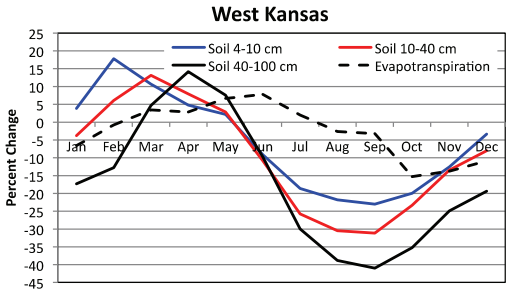



Figure 3. Contours of fluid velocity in a freely evolving polydisperse gas-solid flow.

**Project 4: Coupled Dynamics of Land Use Change and Regional Climate Extremes**

The ISU climate simulation group is developing first-of-its kind agricultural policy-climate projection systems to address food security and climate change[7,8]. The goal is to perform iterative simulations that are both unprecedented in model coupling and in capability for addressing uncertainty growth in these projection systems. Decision makers rely on projections of climate, land use, and agricultural productivity to develop policies that promote increased production and acceleration of efficient agricultural practices in developing countries. Current projection systems are not dynami-

Figure 4. Percent change of soil moisture and evapotranspiration across western Kansas predicted by regional climate simulations. Two 25-yr simulations were generated with identical weather conditions but differing land use. (Change is current land use minus POLYSYS projection).



cally coupled, so that feedbacks between land use and climate change are not considered. We are integrating projections of climate change and policy-driven agricultural land use change through our novel design of climate and agricultural projection systems. For example, projected land use change to increase switchgrass production in the Great Plains creates a soil water deficit that reduces plant transpiration during summer and fall (Figure 4), and, by extension, biomass productivity. This model-based result, in combination with similar findings from hay cropping systems, strongly argues that water resources are insufficient to attain the policy target.

**Project 5: Other Application Areas**

***Prediction and Discovery of Materials: Reverse Engineering of Crystal Structures.*** The prediction of crystal structures from their chemical composition has long been recognized as one of the outstanding challenges in theoretical solid state physics[9] [10]. From the materials design perspective, it is desirable to have a method that requires no prior knowledge or assumptions about the system. Our aim is to develop reliable computational tools that can predict material structures from given chemical compositions in the emerging new area of computational discovery of materials with optimal properties. Algorithms proposed to tackle this challenging problem include simulated annealing, genetic algorithms (GA), and basin or minima hopping. The GA has proved to be a powerful approach to predict material structures using first principles calculations and the knowledge of the chemical composition. It also scales well in terms of throughput achievable by increasing the number of computing nodes.

***Atom Probe Tomography.*** HPC also is critical to establish a new computational paradigm and infrastructure that will enhance 3-D atomic reconstruction for an emerging new instrumentation technology, atom probe tomography (APT). APT is a powerful microscopy tool that enables spatial resolution of hundreds of millions of atoms at the sub-nanoscale. APT is the only instrument capable of mapping the 3-D composition of a metal, semiconductor, or an insulator with atomic resolution. The 3-D reconstruction of this direct space information provides unprecedented capabilities for characterizing materials at the atomic level.

***Design of Large Energy System Design Optimization.*** With increasing demand of energy, power system expansion must be planned while addressing the integration of various different renewal energy sources, satisfying the policy requirements, and accounting for interdependencies among the energy sources and their transformation. We at ISU are addressing these problems under a NSF supported EFRI project[11] [12]. In capacity planning problems, when total demand in the system is greater than the total supply, we need to include commodity generation and capacity expansion capabilities in the model. We have developed a transformation methodology to transform capacity expansion problems into linear network flow problem, which allows solution of such problems using multiple computers.

***Education***. ISU has developed the Nation's largest bioinformatics and computational biology graduate program and is recognized as a major provider of a trained workforce in these critical areas. The program is supported by two NSF IGERT grants and a long-term REU, and serves as a common educational platform for students pursuing research work in biosciences through our various departments, centers and institutes. ISU also received a new IGERT grant in the wind energy area. We will incorporate the algorithms and methods to carry out large scale computations in two advanced classes. These courses will provide high-level training for students and prepare them for research in high performance computation. The state-of-the-art HPC cluster, coupled with our research programs focused on real world problems, will position ISU to better compete for women and underrepresented minority graduate students.

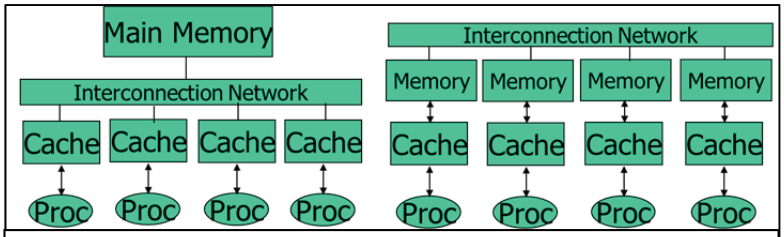


Figure 5: Shared and Distributed Memory Machines

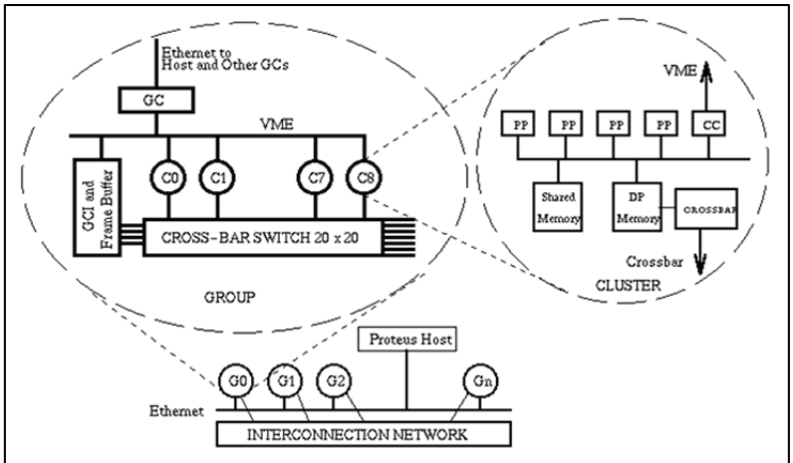


Figure 6: Proteus Distributed Memory Machine with an Efficient and Easy to Use Hierarchical Network

**HPC Infrastructure**

For the desired goals for science, it is clear that a new HPC platform is essential to continue the track record of the extraordinary advances made possible by use of our previous generation machines, the BlueGene/L and Sun cluster systems supporting interdisciplinary and multi-disciplinary research. *Each research project above explicitly defines its need for the new HPC cluster.* Our applications are large and data-driven, and require more memory per node for processing. They generate large volumes of data that must be stored for reuse and sharing. Therefore, we decide to approach the NSF MRI program to propose a cluster configuration that represents a balanced machine and maximizes the use of processing cores, memory, and storage. It was a group effort built upon our culture of sharing resources and our tradition of developing new HPC applications in collaboration with domain experts to solve leading-edge problems in science and engineering.

**Earlier Machines**. In the past (1970s, 1980s, 1990s) when research needed computing, specific machines were designed to fulfill the need[13]. Examples include Maspar, Thinking machine, Intel Paragon, IBM SP1, SP2 etc. They were based on two dominant models (as shown in Figure 5): i) Shared Memory machines where memory is shared by all processors; and ii) Distributed Memory machines where data is exchanged through messages. It is easy to program under the first paradigm, but the machines do not scale. It is bit harder to program under the second paradigm, but machines scale better. In both cases, the main important features for high performance included: i) fast multiple processing elements connected with high-speed networks; ii) efficient partitioning of problem; iii) efficient algorithms using large chunks partitioning and using coarse-grain messages with low overhead per message; and iv) low network latency which is tolerant to communication latency and allowed computation/communication overlap. One example of such a machine is the Proteus machine designed and built by the author at the University of Washington as shown in Figure 6. This machine was used for coarse grain image processing elements and included an efficient computing node and a coarse grain message passing system for large data processing.

**Newer Paradigms**: Earlier machines were difficult to manage and program. Since then much progress has been made in their design and programming. They use commodity processor and networking systems, and are easy to manage/program providing cost to computational efficiency. A new computing paradigm has emerged that includes the following:

1. Infrastructure as a service where rather than buying a machine now, one can rent them from Amazon, GoGrid, AppNexus, HP, and many others. These vendors create a cloud that delivers computing and storage as a service.
2. Platform as a service where users simply require a nice API (application programming interface) and platform designer takes care of the rest of implementation, such as a database, web server, and development tools.
3. Software as a service where users just run their applications like Google email or virtual desktop.
4. A shared infrastructure like cloud that is shared by a large number of users, which is how many universities are structuring their computing infrastructure. These are cost effective if sharing can work.

**HPC Machine**: We adopted the last paradigm for our effort. We started to develop a collaborative HPC infrastructure model with a goal to position ISU strategically for advancement in research. The goals set for our effort included the following:

1. Identify and address the needs of HPC users at ISU campus;
2. Support most users from medium to large scale;
3. Utilize and leverage resources;
4. Achieve economies of scale in space utilization;

5. Maximize the use of physical facilities-power/cooling/racks /cable;
6. Leverage available funds;
7. Longer commitments and unchanged vision;
8. Move from an ad hoc to a planned approach; and
9. Sustainable over long term!

**System Configuration**: When we discussed the needs with our user groups, the following configuration emerged. Our HPC cluster consists of the following types of nodes to meet the anticipated demands.

Head Node. The Head Node is critical to the operation of the entire cluster. Home directories are stored on the head node. It uses hot swappable redundant power supplies, large storage for user directories, and includes a dual port 10Gb NIC with Fiber connections to connect to the campus network.

Three types of Compute Nodes: GPU, Fat, and Thin. Four "Fat" Compute Nodes will have 16GB memory per core and the rest of the Compute Nodes will be "Thin" nodes with 8 GB memory per core. 32 GPU Compute Nodes include dual GPU system.

Interconnects: Compute, Head, and Storage Nodes are interconnected via QDR InfiniBand switch, a Gigabit Ethernet switch and an Ethernet switch for IPMI.

Storage: The storage is designed to be two tiered: 150 TB of fast, reliable storage for scratch space using the Lustre file system plus about 850 TB of raw storage using NFS file system that is scalable, highly reliable, expandable and fault tolerant storage with no single points of failure.

System Software: The system software includes (i) The Red Hat Enterprise Operating System; (ii) The Lustre parallel file system; (iii) Intel's Fortran, C and C++ Compilers; (iv) Intel's Cluster Toolkit; (iv) Intel's MPI; (v) PGI's Fortran, C and C++ Compilers; (vi) PGI's CUDA Fortran; (vii) GNU's Fortran, C and C++ Compilers; (viii) Allinea's DDT and OPT; and (ix) The MOAB's Cluster Suite for the GPUs.

Racks are cooled so that little or no heat is put into the machine room.

**Realization**: A group of faculty members was identified to develop external funding for this proposal, specifically targeting the NSF MRI (major research instrumentation) program. A successful $2.6M proposal for a large heterogeneous machine was developed that met the needs of a plurality of researchers. It is hard to bring people together, in particular from multiple disciplines, but not impossible. But once done, it is worth the effort. We at ISU are very proud of our success.

**Acknowledgements**: The project description is based on details provided by various researchers including Drs. Srinivas Aluru, Patrick Schnabel, Jim Reecy, Robert Jernigan, M. Gordon, Theresa Windus, Monica Lamm, Rodney Fox, Z. J. Wang, Baskar Ganapathy Subramanian, Shankar Subramaniam, Eugene Takle, Chistopher Anderson, Bruce Harmon, Krishna Rajan, Jim McCalley, and Glenn Luecke.

**References**

1. S. Gnerre, I. MacCallum, D. Przybylski, F.J. Riberiro *et al*. "High-quality draft assemblies of mammalian genomes from massively

parallel sequence data", *Proceedings of the National Academy of Sciences USA*, 108, 1513 (2011).
2. J. Zola, M. Aluru, A. Sarje and S. Aluru, "Parallel information theory based construction of gene regulatory networks," *IEEE Transactions on Parallel and Distributed Systems, Vol.* 21 (12), 1721 (2010).
3. Y. Tamada, S. Imoto, H. Araki, M. Nagasaki *et al*. "Estimating genome-wide gene networks using nonparametric Bayesian network models on massively parallel computers", *ACM/IEEE Transactions on Computational Biology and Bioinformatics*, 8, No. 3, pp. 683 (2011).
4. G. Pranami, L. Slipchenko, M. H. Lamm and M S. Gordon, "Coarse-grained intermolecular potentials derived from the effective fragment potential: Application to water, benzene, and carbon tetrachloride," in Multi-scale Quantum Models for Biocatalysis: Modern Techniques and Applications, D.M. York and T.-S. Lee, Eds., Springer-Verlag, 2009.
5. Asadchev, V. Allada, J. Felder, B. Bode, M.S. Gordon, and T.L. Windus, "Uncontracted two-electron repulsion integral implementation on multicores and GPUs using the Rys quadrature method", *J. Chem. Theory Comput.* 6, 696 (2010)
6. S. Tenneti, R. Garg, C. M. Hrenya, R. O. Fox and S. Subramaniam, "Direct numerical simulation of gas-solid suspensions at moderate Reynolds number: Quantifying the coupling between hydrodynamic forces and particle velocity fluctuations," *Powder Technology* 203, 57 (2010).
7. Georgescu, M., D. B. Lobell, and C. B. Field "Direct climate effects of perennial bioenergy crops in the United States," *Pro. Nat. Acad. Sci.*, 108, 4307 (2011).
8. G. Hoogenboom, "Decision support system to study climate change impacts on crop production. Special publication," American Society of Agronomy ASA Special Publication, 2003. Also see http://www.icasa.net/dssat/.
9. J. Maddox, "Crystals from first principles," *Nature*, 335, 201 (1988).
10. K.D.M. Harris, R.L. Johnston, B.M. Kariuki, "The genetic algorithm: Foundations and applications in structure solution from powder diffraction data," *Acta Crystallogr.* A54, 632 (1998)
11. E. Ibanez, and J. McCalley, "Multiobjective evolutionary algorithm for long-term planning of the national energy and transportation systems," Energy Systems Journal, Vol. 2(2), pp. 151-169 (2011).
12. A.K. Somani and Jinxu Ding, "Parallel Computing Solution for Capacity Expansion Network Flow Optimization Problems," Journal of Computing, (July 2012).
13. A.K. Somani, C. Wittenbrink, R. M. Haralick, L. G. Shapiro, J. N. Hwang, C. Chen, R. Johnson, and K. Cooper, "Proteus System Architecture and Organization," in the Proc. of 5th International Parallel Processing Symposium, June 1991, pp. 287-294.

**RETREAT PARTICIPANTS 2012**

## Keynote Speaker

**David Shulenburger, PhD**
Emeritus Vice President for Academic Affairs, Association of Public and Land Grant Universities

## Iowa State University

**James A. Davis, PhD,** Vice Provost for IT and CIO
**Arun K. Somani, PhD,** Anson Marston Distinguished Professor

## Kansas State University

**James Guikema, PhD,** Associate Vice President, Office of Research and Sponsored Programs
**Susan Brown, PhD,** University Distinguished Professor of Biology

## The University of Kansas

**Jeffrey Scott Vitter, PhD,** Provost and Executive Vice Chancellor
**Danny Anderson, PhD,** Dean, College of Liberal Arts & Sciences
**Mabel L. Rice, PhD,** Fred & Virginia Merrill Distinguished Professor of Advanced Studies; Director of the Merrill Center
**Perry Alexander, PhD,** Director, Information and Telecommunication Technology Center (ITTC); Professor, Electrical Engineering and Computer Science
**John Colombo, PhD,** Director, Life Span Institute
**Mary Lee Hummert, PhD,** Vice Provost for Faculty Development
**Deb Teeter,** University Director, Office of Institutional Research and Planning
**Steven Warren, PhD,** Vice Chancellor for Research & Graduate Studies

## University of Kansas Medical Center

**Paul Terranova, PhD,** Vice Chancellor for Research; Senior Associate Dean for Research and Graduate Education
**Lemuel Russell Waitman, PhD,** Director, Medical Informatics

## University of Missouri

**Brian Foster, PhD,** Provost
**Robert V. Duncan, PhD,** Vice Chancellor for Research
**Gary K. Allen, PhD,** Vice President for Information Technology
**Chi-Ren Shyu, PhD,** Director, MU Informatics Institute; Paul K. and Diane Shumaker Professor of Biomedical Informatics.

## University of Nebraska-Lincoln

**Prem S. Paul, PhD,** Vice Chancellor for Research and Economic Development
**David Swanson, PhD,** Director, Holland Computing Center

## Other Participants

**Margaret Echelbarger,** Doctoral Trainee in Child Language, KU