e-Science Task Force Final Report

February 2009

Introduction

Columbia University is one of the pre-eminent research institutions in the country, with sponsored project expenditures of $789 million in FY 2006-2007. The use of technology in all its rapidly changing forms is essential to research—especially as its pursuit and impact becomes increasingly reliant on more powerful computing, enhanced access to data, and more pervasive global communication and collaboration. An infrastructure and strategy that support the use of new technologies are therefore critical to continuing the University's leadership in research, teaching, and learning.

In several recent reports and solicitations, the National Science Foundation (NSF) has not only emphasized the importance of shared technological resources for academic work and research in all disciplines, but also pointed out the need to embed resources, tools, and services within a larger system—a cyberinfrastructure (CI). Based on networks, computers, and data storage, a CI's foundation is the services, software, and human expertise that organize these resources to make them ubiquitously and seamlessly accessible to researchers, faculty, and students. Such a system aims to enable "distributed knowledge communities that collaborate and communicate across disciplines, distances, and cultures."[1]

Echoing NSF, a research initiatives report solicited by the Computing Research Association in December 2008 recommends "a series of investments to create balanced high performance cyberinfrastructure for hundreds of U.S. colleges and universities which will stimulate the development, deployment, and application of a new generation of data-intensive discovery." Because, as the report states, "Research universities are the central engine of the innovation economy," the task of "providing network enabled opportunites for students and faculty to work with large-scale, data-intensive computing and other cyberinfrastructure will yield high returns over many years."[2] Columbia must be in a position to take advantage of such national cyberinfrastructure investments should they occur.

---

[1] "Cyberinfrastructure Vision for 21st Century Discovery." National Science Foundation. http://www.nsf.gov/pubs/2007/nsf0728/index.jsp.

[2] Ed Lazowska, Peter Lee, Chip Elliott, and Larry Smarr. Infrastructure for eScience and eLearning in Higher Education. December 2008. <http://www.cra.org/ccc/docs/init/Infrastructure.pdf>

Elements of CI are already in use by many at Columbia to begin to meet the undeniable need for widely dispersed research teams to share resources and jointly develop hypotheses, experiments, analyses, and publications. Researchers and scholars are building and sharing computer clusters, joining global consortia to access high-performance computers and data on and off campus, and using the collaborative software that ties these resources together. This trend demonstrates that the highly decentralized, individually administered compute cluster model of the past is no longer able to meet the needs of many researchers, who now require a richer and more powerful tool set.

To maintain its place as a leading research institution, Columbia must develop a University strategy for coordinating, supporting, and developing e-science and the CI it depends on in order to ensure and optimize the required investments in technological and human resources.

University Librarian James Neal, with the support of Provost Alan Brinkley and Executive Vice President for Research David Hirsh, convened the e-Science Task Force in March 2008. A Working Group of Task Force members talked with over 50 researchers and administrators across the University—at the Morningside Campus, Columbia University Medical Center (CUMC), and Lamont campus.

Through eight months of interviews and discussions, the Working Group found overwhelming evidence that the Columbia research community's technology and computing support needs are not being adequately met. Faculty and administrators from every division of the University stated that much can be done to improve the resources available, as well as to relieve the present burdens on researchers' time and finances engendered by the existing, decentralized model of support. Recruitment and retention of faculty and graduate students were also cited as a significant problem, with recruits from other universities expecting computing resources to be centrally provided or, finding this is not the case, negotiating aggressively for additional resources in their start-up packages. Current faculty also point with envy to the resources available to their peers at other institutions.

The interviews, discussions, and surveys identified many issues, from which the Task Force distilled seven critical areas. Within these seven areas, we identified four with critical, broad impact that we believe are feasible to undertake in terms of available time and money. To address the current state of research computing at the University and to foster research momentum and potential going forward, the Task Force recommends that these four areas take priority for immediate action:

- High-performance computing
- Data storage and archiving
- Networking
- Governance and policy

The Task Force cannot overemphasize that the future of research at Columbia is dependent on a shift away from the current decentralized model and towards a system based on shared resources, comprehensive support, and a clear CI strategy. The result promises a more cost-effective model if implemented in a coordinated fashion, as well as a crucial framework for sustaining our competitiveness for research and the funding on which it depends. As research dollars become increasingly difficult to secure, the University cannot afford an inadequate CI now that federal agencies are requiring, and our peer institutions are implementing, CI services and support.

In a time of rapid technological change and ever-widening possibilities for the production of new knowledge, our peer research institutions aiming to be at the forefront are equipping their researchers and scholars with the CI necessary to pursue increasingly complex analyses; to produce, manage, share, and preserve larger amounts of data; and to collaborate frequently across disciplines, institutions, and national and international borders.

In the following pages, we detail our recommendations.

High-Performance Computing

High-performance computing (HPC) is vital to the agenda of researchers in physical and social sciences, engineering, medicine, and public health at Columbia. Even departments that are not traditionally associated with HPC, from Statistics to Pharmacology to Political Science, are deploying new research tools and methods and recruiting new faculty and students who require access to HPC environments.

The e-Science Task Force's information-gathering phase confirmed the emergence of these 'mid-tier' researchers whose need for computing resources exceeds the capabilities of a single workstation or small cluster, but falls well short of that of a super-computer. These researchers lack the funding and technical support to maintain an effective technology infrastructure.

At Columbia over the last two decades, technical trends and past federal funding practices, as well as the entrepreneurial climate, encouraged the creation of many small independent clusters. These clusters are located in disparate places throughout the campus, consuming valuable space and energy. Within the clusters, there are often minimal backups performed and variable systems administration experience resulting in inefficiency, more frequent downtimes, and potential loss or corruption of data. Another, and perhaps the most costly, penalty is exacted in the significant loss of research time for faculty and graduate students who must administer these systems.

Many of our peer institutions encourage faculty to participate in a larger, centrally administered cluster as a way of using cost-sharing to capture economies of scale and more effectively deploying capital and operating funds. The rising price of energy and green concerns are accentuating this trend.[3]

We recommend that Columbia:

(1) Support the development of a centrally located and administered 64-128 node cluster to form the core of a general HPC service. A nucleus of equipment and three years of operational support should be underwritten to encourage movement of dispersed clusters to the data center.
(2) Develop an intermediate term plan to provide service to the new occupants of the Interdisciplinary Science Building (ISB), which will have no server rooms, to conserve this new, expensive lab space for research.
(3) Develop a long-term plan for HPC services that considers the total cost of ownership for the University and rationalizes practices and investments to maximize benefit.

The estimated annual cost for item (1) is $150,000 to $200,000 including equipment, amortization, and staffing for the initial three years.

The benefits of providing hosting space in a central, professionally managed data center for school, department, and project clusters would return faculty and students to their primary responsibilities of research, teaching, and learning rather than spending time as system administrators. Departments could more easily and economically craft start-up packages. The University as a whole would save substantially on one-time renovation costs for clusters as well as annual energy and air-conditioning costs. The full systems administration and support services available in such shared clusters would provide significantly enhanced security. Further, central HPC facilities would allow researchers without clusters to explore such systems for developing proposals and securing research grants. At the same time, such facilities would also permit greatly expanded use of HPC techniques in instruction for those who wish to employ them.

[3] "IT Engagement in Research: A Baseline Study" (2006); "IT Engagement in Research: A View of Medical School Practice" (2008). Educause. http://www.educause.edu/.

Data Storage and Archiving

In contrast to the past, when research data were held in filing cabinets or in lab notebooks in physical form, today's research outputs often unwittingly become electronic ephemera. Digital products of research, frequently stored only on local hard drives or servers, are vulnerable to overwriting or deletion to make way for more recent materials, and are often lost due to software changes or hardware failures.

Granting agency requirements for data sharing have also made scholarly communication of research more complex. The National Institutes of Health (NIH) and the National Science Foundation (NSF) both require data sharing as part of their grant funding, but neither agency makes provision for fulfilling this requirement. (This is in contrast with NASA and NOAA, which provide significant and systematic support for maintaining accessible digital data collections through distributed active archive centers.) Other funding agencies are likely to follow NIH and NSF's lead. The Federal Interagency Working Group on Digital Data was established in 2007 with the eventual goal that data resulting from any federally funded grants must be made publicly available. Several reports have shown that local support is critical for the success and regulatory compliance of data sharing efforts,[4] yet currently Columbia has no University policies or services in place to support researcher-generated data collections, such as those defined by the National Science Board.[5] Many of our peer institutions have plans for putting this support in place, if they have not done so already.[6] While some work has been done in this area through a Long Term Archive (LTA) pilot project, this LTA has been designed to support long-term data and information stewardship solely for a large "reference" collection, the NASA Socioeconomic Data and Applications Center (SEDAC) operated by CIESIN, a research unit of the Earth Institute.[7]

---

[4] Glover, D. M., C. L. Chandler, S. C. Doney, K. O. Buesseler, G. Heimerdinger, J. K. B. Bishop, and G. R. Flierl, "The US JGOFS data management experience." *Deep-Sea Research Part II: Topical Studies in Oceanography* 53(5-7) (2006): 793-802.

Karasti, H., K. S. Baker, and E. Halkola, "Enriching the notion of data curation in e-science: Data managing and information infrastructuring in the long term ecological research (LTER) network." *Computer Supported Cooperative Work: CSCW: An International Journal* 15, (4) (2006): 321-58.

Lord, P., and A. Macdonald, "e-Science curation report data curation for e-science in the UK: An audit to establish requirements for future curation and provision." JISC Committee for the Support of Research (2003).

[5] "Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century," National Science Board publication NSB-05-40. http://www.nsf.gov/pubs/2005/nsb0540/.

[6] See, for example, the report of the Cornell University Library Data Working Group entitled "Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities for the Cornell University Library" (May 2008).

[7] "About the SEDAC Long-Term Archive." SEDAC Socioeconomic Data and Application Center. http://sedac.ciesin.columbia.edu/lta/About_SEDAC_LTA.html.

While several Columbia research groups have obtained direct funding to support data storage for their research output, some have expressed concern that these solutions are not sustainable and rely on temporary grant or agency funding with no long-term strategy. Other Columbia researchers with less funding or expertise are unable to manage their data collections with adequate access, backup, documentation, curation, or migration. This situation risks data loss, which would eliminate future access to the scientific record.

The Columbia Libraries/Information Services' (CUL/IS) Academic Commons repository has been set up to collect, preserve, and make accessible through search and discovery tools the scholarship of the faculty. This research output may include datasets and raw data—the objects required for deposit by funding agencies—as well as materials that help to contextualize that data, such as articles, book chapters, essays, monographs, working papers, technical reports, conference presentations, multimedia creations (e.g., simulations, three-dimensional maps), and other materials in digital formats.

In order to begin understanding the scope of the problem and planning for serving the needs of the University in future years, we recommend that Columbia:

(1) Prepare for strengthened NIH and NSF data sharing mandates by conducting a survey, to be directed by the Office of the Executive Vice President for Research and CUL/IS, of research data storage needs of funded NIH and NSF Columbia research grants from 2009 onward.
(2) Based on the survey results, expand the existing CUL/IS Academic Commons to collect, preserve, and make accessible the data currently required by NIH and NSF funding (and by other funding agencies), for which there is not otherwise an established data archive.
(3) Develop a long-term plan to expand the Academic Commons to fully support federal agency requirements for long-term data preservation and access.

We estimate the annual operating cost of supplementing the repository with an additional 40 Terabytes of storage to be $250,000 to $300,000 including equipment, amortization, and staffing.

The content placed in the Academic Commons can become part of a global interoperable repository system that allows for data sharing of related research results from around the world. While our present capacity is not adequate for significant amounts of numeric and spatial data storage, our recommended expansion will allow for both fulfilling NIH and NSF data sharing mandates for a select but substantial number of research groups identified in the joint survey, and for supporting plans by CUL/IS to store more widely the digital data collections deposited by Columbia researchers. This will also serve as the foundation for efforts to obtain external funding support from NSF, NIH, and other agencies for expanded CI capabilities.

Networking

e-Science activities rely on fast, dependable computing networks. The data network on the Morningside campus, while having a high-speed backbone and good connectivity to regional, national, and international research networks, suffers from deferred maintenance at the building level. Network electronics, which should normally be replaced every 4-5 years, are 10 years old in many areas. The Ethernet wiring in about 40% of campus is 20 years old and at 10 megabits per second (Mbps) cannot support current standard speeds of 100 Mbps and 1 gigabit per second (Gbps) connections. Wireless (WiFi) networking coverage is spotty, with only about 25% of the campus covered by robust, centrally managed infrastructure, and the large number of uncoordinated "volunteer" wireless networks on campus create security vulnerabilities. Furthermore, network connectivity to our two significant research campuses, Lamont and Nevis Laboratories, is inadequate. Network capability at the CUMC campus is satisfactory. CUIT is working with Facilities on all Manhattanville construction projects to ensure that advanced networking capabilities are available.

Columbia University Information Technology (CUIT) is developing a network upgrade proposal, for a FY2010 funding request, that addresses replacing the network electronics to establish newer, more secure, higher performance networking, as well as deploying ubiquitous high-performance (IEEE 802.11n) WiFi campus-wide.

The Task Force recommends that research facilities most affected by inadequate networking receive immediate attention. To accomplish this goal, CUIT will propose that a fund pool and governance body be established to assist in implementing the necessary renewals.

The estimated cost for the portion of the proposal that directly applies to research labs, offices, and the underserved Lamont and Nevis campuses is a one-time amount of $1.5-$3M with approximately $500,000-$700,000 in annual costs.

The Columbia network upgrade project will address insufficient network capacity in certain locations and give researchers the capability to work with very large datasets. Evenly distributed wireless coverage will allow use of a single laptop between laboratory, office, and home, and accommodate University visitors, including research colleagues from elsewhere. The upgrade will also allow Columbia to implement a more flexible network access control system where needed.

Governance and Policy

While technology issues drive researchers' concerns about the state of research computing at Columbia, better governance is a basic requirement for making and keeping our research CI competitive. A formal governance structure is the foundation of any sustainable e-science program and its supporting community. Similar efforts at other institutions align research, information technology, and administrative organizations, and implement policies and procedures to identify and then meet researchers' needs.[8]

CI governance must be researcher-driven and accommodate the differing interests and needs of the whole variety of research disciplines. To this end, the governance process must include open, active communication channels that bring researchers' needs to the fore, take advantage of the reservoir of expertise now dispersed across campuses, and encourage the growth of a research technology community for information exchange and mutual support.

Much of Columbia's current research eminence is derived from the active entrepreneurial spirit of our researchers in obtaining and using the latest technology, which has, in turn, helped drive dramatic research successes. But the absence of a University focus on research technology has resulted in inefficiencies arising from the dozens, perhaps hundreds, of independent computing facilities now scattered in departments and labs. The federal agencies have clearly signaled a diminishing enthusiasm for supporting individual clusters. Therefore, the governance process must seek out and reduce inefficiencies, and provide mechanisms for investing the savings to better support the research enterprise.

The current semi-formal or ad hoc methods of coordination, where they exist, are inadequate. Therefore, the Task Force recommends that Columbia:

(1) Establish a Research Technology Council responsible to the Provost, the Executive Vice President for Research, and the Senior Executive Vice President.
(2) Include on the Council faculty from each major disciplinary area and representatives from each campus' central IT unit and appropriate administrative offices, such as the Provost, Research, Libraries, Facilities, Finance, Environmental Stewardship, and Student and Administrative Services.
(3) Charge the Council with the responsibility and authority to
    a. Develop a long-term strategic plan for research technology across the University, including coordination of investments at the new Manhattanville campus.
    b. Set policies with respect to the funding, acquisition, use, and delivery of research technology and services for research; and

[8] "IT Engagement in Research: A Baseline Study" (2006); "Process and Politics: IT Governance in Higher Education," Educause. http://www.educause.edu/.

c. Monitor the delivery of such services.

(4) Embed the governance of research technology in the University process for the governance of information technology, generally.

A formal, active governance process for research technology will allow the University to set priorities and focus its limited resources on accomplishing them.

Other Critical Issues

In addition to the four main priorities of high-performance computing, data storage and archiving, networking, and governance, the Task Force proposes that the future governance group address the following critical areas identified through interviews, discussions, and surveys.

Sustainable Funding; Grant Support Improvements

Columbia's current decentralized model of research computing limits the University's ability to understand the real total costs of e-science and e-research, as well as sources of funding. We recommend analyzing funding related to e-research and identifying costs embedded or hidden in various budgets through the following tasks:

(1) Inventory e-research facilities and study their overall institutional costs.
(2) Review charge-back models for detrimental, unintended consequences, e.g., the Morningside $6 jack charge.
(3) Review impact of granting agency changes in support for infrastructure costs.

The Task Force also recommends developing procedures for more flexible use of grant and ICR funding, including sharing of costs among projects and departments. Because there is no budgetary mechanism to identify and recover the savings generated by centralized services and then apply them to these services, we suggest developing methods for applying cost savings from economies of sharing computing resources to support e-research services, e.g., energy savings.

To smooth the process for proposal preparation, we recommend continuing the development of materials and services to support grant proposals such as templates, sample text, proposal libraries, and proposal consultation. We suggest investigating enhancements in RASCAL and InfoEd to ease use of prepared supporting materials.

Coordinated Planning and Procurement; Software Licensing

To assist schools, departments, and researchers in selecting and obtaining technology, the Task Force recommends establishing a research IT consulting and planning service integrated with a coordinated procurement service.

In particular, the absence of a coordinated software-licensing program compels individual researchers to spend excessive amounts of time, effort, and money to obtain software, and increases total University expenditures. We recommend prioritizing software site licensing support to better accommodate research needs.

Domain Consulting; Collaboration Services

New strategies, tools, and instruments developed in different disciplines require expertise that crosses the boundary between the discipline and the technology. The Task Force recommends expanding discipline- and domain-specific consulting being developed by central IT services and CUL/IS.

The absence of University services supporting collaboration in an increasingly global, team-oriented, and cross-disciplinary research environment handicaps our researchers and research proposals. We suggest continuing development of online collaborative services and support by CUL/IS and central IT services to facilitate joint projects on and off campus.

Conclusion

Without a concentrated effort, Columbia may well slip behind other institutions that have already begun to develop new organizational, funding, and governance models to create and support innovative research CI.

Columbia's present research environment offers significant opportunities for development and growth of a sustainable CI. Building such a system and strategy will not only support e-science, but also drive e-research momentum in the social sciences, humanities, and beyond—allowing scholars regardless of field "to focus their intellectual and scholarly energies on the issues that engage them, and to be effective users of new media and new technologies, rather than having to invent them."[9]

The future of Columbia's research impact, as well as the University's ability to continue recruiting and retaining the best scholars, hinges on the development of its CI.

---

[9] "ACLS Commission on Cyberinfrastructure." American Council of Learned Societies. http://www.acls.org/programs/Default.aspx?id=644.

e-Science Task Force Staff

Suzanne Bakken
Alumni Professor of the School of Nursing and Professor of Biomedical Informatics
sbh22@columbia.edu

Peter Bearman
Jonathan R. Cole Professor of Sociology
psb17@columbia.edu

Benno Blumenthal
Data Library Manager, Climate Monitoring and Dissemination, International Research Institute for Climate and Society
benno@iri.columbia.edu

*Walter Bourne
Director, Technology Initiatives (Retired), CUL/IS
walter@columbia.edu

Andrea Califano
Professor of Biomedical Informatics and Chief, Division of Bioinformatics
ac2248@columbia.edu

Lynn Caporale
Associate Director, Judith P. Sulzberger MD Columbia Genome Center
lc2201@columbia.edu

Suzanne Carbotte
Bruce C Heezen Senior Research Scientist in the Lamont-Doherty Earth Observatory
carbotte@ldeo.columbia.edu

*Robert Cartolano
Director, Library Information Technology Office, CUL/IS
rtc@columbia.edu

Robert Chen
Director, Center for International Earth Science Information Network (CIESIN), The Earth Institute
bchen@ciesin.columbia.edu

Richard Clarida
C. Lowell Harriss Professor of Economics and Professor of International and Public Affairs
rhc2@columbia.edu

*Alan Crosswell
Associate Vice President and Chief Technologist, CUIT
alan@columbia.edu

Candace Fleming
Information Technology Vice President and Chief Information Officer, CUIT
cfleming@columbia.edu

Andrew Gelman
Professor, Department of Statistics
gelman@stat.columbia.edu

*Victoria Hamilton
Coordinator, Research Initiatives, Office of the Executive Vice President for Research
victoria.hamilton@columbia.edu

George Hripcsak
Chairman, Department of Biomedical Informatics
gh13@columbia.edu

Emlyn Hughes
Professor, Department of Physics
ewh42@columbia.edu

John Kender
Professor, Department of Computer Science
jrk@cs.columbia.edu

*Rebecca Kennison
Director, Center for Digital Research and Scholarship, CUL/IS
rrk2124@columbia.edu

David Keyes
Fu Foundation Professor, Applied Physics & Applied Mathematics
kd2112@columbia.edu

Ann McDermott
Esther Breslow Professor of Biological Chemistry; Associate Vice President for Academic Planning and Science Initiatives
aem5@columbia.edu

Nilda Mesa
Assistant Vice President of Environmental Stewardship
nm2337@columbia.edu

Shahid Naeem
Chair, Department of Ecology, Evolution and Environmental Biology
sn2121@columbia.edu

James Neal
Vice President for Information Services; University Librarian
jneal@columbia.edu

David Reichman
Professor, Department of Chemistry
drr2103@columbia.edu

*Patricia Renfro
Deputy University Librarian and Associate Vice President for Digital Programs and Technology Services, CUL/IS
pr339@columbia.edu

Peter Schlosser
Vinton Professor of Earth and Environmental Engineering and Professor of Earth and Environmental Sciences; Associate Director, Earth Institute
peters@ldeo.columbia.edu

Robert Shapiro
Professor of Political Science; Director of the Institute for Social and Economic Research and Policy (ISERP)
rys3@columbia.edu

Robert Sideli
Chief Information Officer, Columbia University Medical Center, Assistant Clinical Professor, Biomedical Informatics
rvs1@columbia.edu

John Zimmerman
Associate Professor of Biomedical Informatics; Associate Director, Columbia Center for New Media Teaching and Learning, Assistant Dean for Information Resources, College of Dental Medicine; Associate Professor of Clinical Dentistry
jlz4@columbia.edu

Allen Zweben
Associate Dean for Research and Academic Affairs, School of Social Work
az173@columbia.edu

*Working Group Members