AN IDIOSYNCRATIC PERSPECTIVE ON THE HISTORY AND DEVELOPMENT AT THE UNIVERSITY OF CALIFORNIA, SAN DIEGO OF SUPPORT FOR CYBERINFRASTRUCTURE-ENABLED E-SCIENCE

BRIAN E. C. SCHOTTLAENDER
THE AUDREY GEISEL UNIVERSITY LIBRARIAN
UC SAN DIEGO

ON TERMINOLOGY

I have heard it said that "E-Science" and "Cyberinfrastructure" are, or can be, used interchangeably. I have likewise heard it said that "E-Science" is used in Europe, whereas "Cyberinfrastructure" is used in the U.S.; and, again, they are used to mean the same thing. In this paper, they will not—be used interchangeably or to mean the same thing, that is.

In *Wikipedia* "Cyberinfrastructure" is described as being:

> … the new research environments that support advanced data acquisition, data storage, data management, data integration, data mining, data visualization and other computing and information processing services over the Internet. In scientific usage, cyberinfrastructure is a technological solution to the problem of efficiently connecting data, computers, and people with the goal of enabling derivation of novel scientific theories and knowledge.[1]

Wikipedia also quotes Dr. Francine Berman, Director of the San Diego Supercomputer Center, as describing "cyberinfrastructure" to be :

> … the coordinated aggregate of software, hardware and other technologies, as well as human expertise, required to support current and future discoveries in science and engineering. The challenge of Cyberinfrastructure is to integrate

[1] http://en.wikipedia.org/wiki/Cyberinfrastructure

> relevant and often disparate resources to provide a useful, usable, and enabling framework for research and discovery characterized by broad access and "end-to-end" coordination.[2]

Ironically, *Wikipedia* also observes that "Cyberinfrastructure is also called e-Science …"[3]

"E-science," on the other hand, is described in *Wikipedia* as being:

> … computationally intensive science that is carried out in highly distributed network environments, or science that uses immense data sets that require grid computing; the term sometimes includes technologies that enable distributed collaboration …" [4]

Here too, it is noted that, in the U.S., "… the term cyberinfrastructure is typically used to define e-Science projects …"[5] It is easy enough to understand—what with embedded phrases like "distributed network environments," "grid computing," and "technologies that enable distributed collaboration"—why "E-Science" and "Cyberinfrastructure" are used interchangeably, if not confused. To do so, however, ignores the critical enabling relationship between the two concepts: that is, "Cyberinfrastructure" is the means to the end that is "E-Science." The latter cannot, quite literally, be carried out without the former; and without the latter, the former is little more than a solution in search of a problem. Thus, in this paper, when I am speaking of the means, I'll use "cyberinfrastructure;" when of the end, "E-science."

---

[2] Ibid.
[3] Ibid.
[4] http://en.wikipedia.org/wiki/EScience
[5] Ibid.

EARLY HISTORY: SDSC, 1985–2004

Until recently, the primary manifestation of cyberinfrastructure support for E-Science at UC San Diego was the San Diego Supercomputer Center (SDSC) which was created jointly in 1985 by UCSD and General Atomics, with, primarily, funding from the U.S. National Science Foundation (NSF). Located on the UCSD campus, SDSC operated as a division of General Atomics until 1997, with a research focus on earthquakes, global climate, and other scientific issues requiring massive compute power. When NSF's Supercomputer Center Program came to an end in 1997, SDSC—whose computers had an aggregate capacity of 270 gigaflops, as compared to the 25-30 gigaflop capacity of other computing clusters on the UCSD campus[6]—became an Organized Research Unit (ORU) of UCSD.

In 1996 NSF launched the PACI Program (Partnerships for Advanced Computational Infrastructure) and in 1997 SDSC was named one of two national "leading edge sites" (the other was the University of Illinois' NCSA, the National Center for Supercomputing Applications, led by Dr. Larry Smarr). In addition to compute-intensive hardware, the focus of the PACI program was on partnership, community building, and integrative software.[7]

The cyberinfrastructure assembled at SDSC over the course of this virtually unprecedented 20 years of growth included as of 2005:

- 400 professionals with expertise across multiple scientific domains and computing technologies;
- DataStar, initially a 10.4 teraflops supercomputer ranked amongst the 25 top supercomputers in the world;

[6] "San Diego Supercomputer Center ORU 5-Year Review." Unpublished paper (2005): p. 3.
[7] Ibid.

- An IBM Blue Gene eServer architecture that combines high processor performance with low power consumption;
- Six petabytes of archival tape storage and 500 terabytes of online disk storage; and
- locally developed data management software, including the Storage Resource Broker (SRB).[8]

The E-Science supported by this constellation of resources includes:

- AFCS (Alliance for Cell Signaling); [9]
- BIRN (Biomedical Informatics Research Network);[10]
- CAIDA (Cooperative Association for Internet Data Analysis);[11]
- CiPRES (Cyberinfrastructure for Phylogenetic Research);[12]
- GEON (Geosciences Network);[13]
- HPWREN (High Performance Wireless Research and Education Network);[14]
- NEES (Network for Earthquake Engineering Simulation);[15]
- NEON (National Ecological Observatory Network);[16]
- NLANR (National Laboratory for Applied Network Research);[17]
- ORION (Ocean Research Interactive Observatory Networks; now OOI (Ocean Observatories Initiative));[18]
- PDB (Protein Data Bank);[19]
- PLANT (Plant Genome Research Program);[20]

[8] "San Diego Supercomputer Center ORU 5-Year Review." Unpublished paper (2005): p. 2.
[9] http://www.afcs.org/
[10] http://www.nbirn.net/
[11] http://www.caida.org/home/
[12] http://www.phylo.org/
[13] http://www.geongrid.org/.
[14] http://hpwren.ucsd.edu/
[15] http://www.nees.org/
[16] http://www.neoninc.org/
[17] http://www.nlanr.net/
[18] http://www.oceanleadership.org/ocean_observing
[19] http://www.rcsb.org/pdb/home/home.do

- ROADNet (Real-time Observatories, Applications, and Data Management Network); and
- SEEK (Science Environment for Ecological Knowledge).[21]

MODERN HISTORY (PART I): SDSC, 2005–PRESENT

In 2004 the PACI program was terminated and NSF, instead, provided SDSC and NCSA sufficient "core funding" to continue them as Cyberinfrastructure Centers through 2007, the intended conclusion of PACI. Amongst the differences between the formal (in NSF terms) "Cyberinfrastructure Era" that is now upon us and the data science ecosystem that preceded it is that "enabling technologies must be coordinated, integrated, usable, and interoperable"[22] in order to meet the varying needs of multiple domains and user communities. In addition, the present ecosystem includes players additional to the supercomputer centers, PIs (Private Investigators), and federal funding agencies who had previously populated it, including libraries and university academic and administrative computing centers. Moreover, the present environment is suffused with far greater anxiety about, and consequent attention to, data access issues, both in the public policy arena and the preservation and archiving arena. Finally, there are far fewer financial resources available presently, or at least far greater competition for those resources. All of these factors have had, and will have, implications for SDSC and for the shape cyberinfrastructure support at UCSD is taking, the ends to which that infrastructure will be put, and the financial and organizational models likely needed to achieve those ends.

Between the succession of Dr. Francine Berman to the SDSC Directorship in 2001and the demise of the PACI Program in 2004, SDSC focused increasingly on establishing collaborative relationships

[20] http://usinfo.state.gov/ei/Archive/2004/Jan/28-331488.html
[21] http://seek.ecoinformatics.org/
[22] "San Diego Supercomputer Center ORU 5-Year Review." Unpublished paper (2005): p. 4.

on the UCSD campus, including partnerships with the UCSD Libraries and the California Institute of Information Technology and Telecommunications (Calit2, directed by Dr. Larry Smarr, formerly of NCSA). Both the Libraries and Calit2 are in the information business, if not the knowledge business, and SDSC's partnership overtures make clear that it has realized that it is no longer (just) in the computing business, but in the same information business as well.

What, in fact, the three campus agencies have in common is data. In the 2005 ORU Review of SDSC, Fran Berman noted that during the period 2006-2010 the Center would provide "Professional-level data services, software, and curation beyond what is feasible in university/campus/research lab facilities, and not available in the private sector," adding that "This focus is also helping us diversify SDSC's funding portfolio, adding data-oriented projects and funding from NIH, the Library of Congress, the National Archives, the museum and library communities, and others, in addition to our traditional support from NSF."[23]

Campus and System Context

In the last several years, there have been a variety of conversations on- and off-campus that have helped shape UC San Diego's commitment to supporting E-Science. In 2002, then-UCSD Chancellor (now-UC President) Robert Dynes constituted and charged a campuswide Technology Directions Committee, chaired by Dr. Sid Karin, former (and founding) Director of the San Diego Supercomputer Center. During its three years of existence, the committee explored a variety of technology-related themes.

[23] "San Diego Supercomputer Center ORU 5-Year Review." Unpublished paper (2005): p. 6.

I chaired the subcommittee that led the group's deliberations in regards to content. The following questions informed our thinking:

- What intellectual assets do we have?
- How do/could they add value to what we do/want to do?
- What facilitates/would facilitate their adding that value?[24]

In colloquial terms, those questions were in turn answered as follows:

- Place(s) to put them.
- Tools for getting them there.
- Standards for encoding/describing them.
- Tools for discovering them.
- Tools for manipulating them.
- Tools for sharing them.
- Tools for protecting them.[25]

The group concluded that the infrastructure necessary to satisfy the foregoing "requirements" would need to include the following:

- Institutional repository(ies);
- Preservation repository(ies);
- Metadata registry(ies);
- Discovery interfaces;
- Security/Use protocol; and.
- Migration/preservation strategies.[26]

[24] Author's presentation to the UCSD Technology Directions Committee, November 2002.
[25] Ibid.
[26] Ibid.

All of these cyberinfrastructure components are now either developed or in development.

When current UCSD Chancellor Marye Anne Fox arrived in 2005, she instituted an annual series of senior management retreats intended to foster discussion of the most pressing issues facing the campus, data stewardship amongst them. In this context, the use not of the word "management" ("the conducting or supervising of something (as a business)"[27]), but rather of the word "stewardship" ("the careful and responsible management of something entrusted to one's care"[28]) is significant for two reasons. First, rhetorically, "stewardship" suggests, as it is intended to, greater attention or even empathy than does "management:" careful, responsible, entrusted. Second, and less obviously, it signifies the campus' growing recognition that there is more to data curation than simple bit management—that, indeed, data curation is rather more in the nature of a custodial suite of services and policies than it is a managerial one.

At the UC San Diego Chancellor's Retreat of 2006, Fran Berman and I were asked to articulate the needs of the campus with regard to data stewardship. We noted that these assets are increasingly at risk, whether as a consequence of resource scarcity, technology evolution (including storage and delivery systems, access mechanisms, and encoding formats), calamity, or inaction. We went on to describe a then-unique partnership (now, less so) between the San Diego Supercomputer Center and the UCSD Libraries—a partnership planned and poised to mitigate the risks enumerated above. More on that partnership shortly.

[27] *Merriam-Webster's Collegiate Dictionary*. 10th ed. (2002): p. 704.
[28] *Merriam-Webster's Collegiate Dictionary*. 10th ed. (2002): p. 1150.

In parallel with these campus-level discussions, in 2006 University of California Provost Dr. Wyatt R. (Rory) Hume charged a systemwide Information Technology Guidance Committee (ITGC) "to engage in a consultative, 18-month systemwide planning process to … :

- Identify strategic directions for IT investments that enable campuses to meet their distinctive needs more effectively while supporting the University's broader mission, academic programs and strategic goals.
- Promote the deployment of information technology services to support innovation and the enhancement of academic quality and institutional competitiveness.
- Leverage IT investment and expertise to fully exploit collective and campus-specific IT capabilities."[29]

The ITGC had six focus areas, each with an Expert Working Group:

- Advanced Networking Services;
- Common IT Architecture;
- High Performance Research Computing;
- Instructional Technology;
- IT in Student Experience; and
- Stewardship of Digital Assets (which I chaired).

Tellingly, the final report of the ITGC, submitted in December 2007, is titled: "Creating a UC Cyberinfrastructure." The report states unequivocally that "Development of a University of California[-wide] cyberinfrastructure is critical to our success …"[30] The report contains only nine recommendations, a seemingly modest number:

[29] http://www.universityofcalifornia.edu/itgc/charge/welcome.html
[30] http://www.universityofcalifornia.edu/itgc/ITGC_final%20report_bw.pdf: p. 8.

1. Establish the IT Leadership Council as the UC-wide IT governing body.
   "… in close collaboration with academic and administrative leaders at both the campus and systemwide levels."

2. Fund IT as critical infrastructure.
   "… change current funding models to provide sustainable, renewable funding ..."

3. Apply proven collaboration models.
   "Collaboration is the way forward."

4. Invest in network connectivity.
   "… by continually expanding network bandwidth and computing capabilities …"

5. Plan for the next-generation UC data center infrastructure.
   "… develop a new blueprint for providing scalable data center services to the UC community."

6. Develop IT infrastructure, tools, and services to support collaboration within the UC community.

7. Develop UC Grid research cyberinfrastructure services.
   "… deliver reliable, robust high-performance computing services and tools to research faculty who do not need (or cannot afford) to manage their own …"

8. Create the capacity to manage our digital assets.
   "… by adopting strategies to ensure that the information produced in the course of research and instruction is effectively secured, managed, preserved, and made available for appropriate use …"

9. Cultivate organizational leadership for instructional technology and IT in the student experience.
   "… providing learners with enhanced and new IT-enabled educational opportunities."[31]

Even a cursory reading, however, reveals these to be a much taller order than their relatively sparse number might suggest.

MODERN HISTORY (PART II): SDSC AND UCSDL

The mission of SDSC is to "innovate, develop, and use technology to advance science,"[32] while that of the UCSD Libraries (UCSDL) is to "be leaders in providing and promoting information

[31] http://www.universityofcalifornia.edu/itgc/supdocs/ITGC_1pg080317.pdf
[32] http://www.calit2.net/newsroom/article.php?id=307

resources and services to the UCSD community when, where, and how users want them."[33] What is it that these two missions have in common? Each implies the need for substantial curation and preservation infrastructure. As the information environment in which each organization operates has become increasingly data-rich, it has likewise become increasingly obvious that leveraging investments made by each will redound not only to the benefit of both but to that of the larger organization within which we both operate, namely UCSD.

SDSC and UCSDL began partnering on data-intensive projects in 2001 when, joined by the Scripps Institution of Oceanography, we successfully competed for one of the NSDL I grants (National Science Digital Library, Round I). We repeated that in 2003 when, again with Scripps, we successfully competed for an NSDL II grant (National Science Digital Library, Round II). Both of those initiatives called for the development of tools and services that facilitate ingest and management of data and publications arising from, or incorporating, those data. Both also involved the build-out of the "Storage Resource Broker" (SRB), SDSC-developed middleware that "supports shared collections that can be distributed across multiple organizations and heterogeneous storage systems."[34] SRB has several distinguishing features, including: 1) it is agnostic as to content type and format and 2) it "containerizes" data. This latter feature is a particularly compelling one because it allows various curatorial activities (e.g., ingest, organization, discovery, access control) to be carried out at the container level, reducing in the process the various overheads associated with having to do so more atomically.

[33] http://www.ucsd.edu/portal/site/Libraries/menuitem.346352c02aac0c82b9ba4310d34b01ca/?vgnextoid=238a2b3401904110VgnVCM10000045b410acRCRD
[34] http://www.sdsc.edu/srb/index.php/Main_Page

Since entering into those initiatives, SDSC has collaborated ever more closely and deliberately with not only the UCSD Libraries, but other libraries as well at the regional and national levels. At the local level, the UCSD Libraries have used an SRB instance as the basis for developing our own Digital Asset Management System (DAMS). Once an object collection is identified, the Libraries' Metadata Analysis and Specification Unit creates an assembly plan; maps data to MODS, PREMIS, MIX, or various local schemas as necessary; and ingests the collection into the SRB. Original digital objects, their technical metadata, and their descriptive metadata are all stored in the SRB. The UCSDL DAMS currently has 6 Tb of content under active management, including texts, images (still and moving), and sound files. The Libraries have begun discussions with campus academic colleagues about bringing data archives under curatorial management as well.

At the UC-systemwide level, SDSC and the California Digital Library (CDL) are partnering to manage the CDL's Digital Preservation Repository. In addition to providing back-end storage for the DPR, SDSC is collaborating with CDL on the so-called "Mass Transit" project, an initiative to "better understand issues in large-scale transfer and replication of data in the context of digital preservation … [whose] … primary deliverables will include extensive data transfer tests and a publicly available, jointly authored CDL/SDSC document on best practices and situation-based recommendations for institutions embarking on large data transfer in the context of preservation."[35] The DPR is currently scaled to accommodate 40 Tb of content under active management.

This latter project is but one example of UCSD's growing leadership in the arena of long-term digital data preservation. Chronopolis—a collaborative project between SDSC, the UCSD Libraries,

[35] http://masstransit.sdsc.edu/

the U.S. National Center for Atmospheric Research (NCAR), and the University of Maryland—is another.

THE PRESERVATION IMPERATIVE

Long-Lived Digital Data Collections …, published by the U.S. National Science Board in 2005, stated that "Long-lived digital data collections are powerful catalysts for progress and for democratization of science and education."[36] It defined data as:

> … any information that can be stored in digital form, including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc. Such data may be generated by various means including observation, computation, or experiment ...[37]

and went on to describe "long-lived" as extending for a period of time "long enough for there to be concern about the impacts of changing technology."[38] Finally, while this concluding exhortation in the report is directed specifically to the National Science Foundation, it represents a call to action for us all:

> Given the proliferation of resource and reference collections and the costs associated with creating and maintaining them, it is imperative that the Foundation develop a comprehensive strategy – incorporating and integrating technical and financial considerations – for long-lived data collections and determine the steps necessary to anticipate future needs.[39]

---

[36] http://www.nsf.gov/pubs/2005/nsb0540/: p. 9.
[37] Ibid.
[38] Ibid.
[39] http://www.nsf.gov/pubs/2005/nsb0540/: p. 44.

UC San Diego has not only heeded that call, we—and our partners—may well have had a hand in prompting it.

In 2004, SDSC, UCSDL, NCAR, and the University of Maryland, did something highly unusual: we submitted an unsolicited proposal to the U.S. National Science Foundation, titled "Chronopolis: Federated Digital Preservation Across Time and Space." The proposal's project summary read as follows::

> There is a critical need to organize, preserve, and make accessible the increasing number of digital holdings that represent intellectual capital. This intellectual capital contains scientific records that are the basis for current research, future scientific advances, and education source materials for use by the public, educators, scientists and engineers now and for the foreseeable future. We propose a national center for the management, long-term preservation, and promulgation of national digital assets, Chronopolis.
>
> Chronopolis will provide a model facility that enables long-term support of irreplaceable and important national data collections, ensuring that: 1) Standard reference datasets remain available to provide critical science reference material; 2) Collections can expand and evolve over time, as well as weather evolution in the underlying technologies; and 3) Preservation "of last resort" is available for critical disciplinary and interdisciplinary digital resources at risk of being lost.
>
> Chronopolis will provide tools, software, and services needed to manage data, information, and knowledge at the scales required for national digital holdings. It

> will function as a distributed national "data backbone," federating data and information (preservation over "space"), and will provide operational data services for maintaining key digital collections for the long term, ranging from scientific databases to library holdings (preservation over "time"). Chronopolis will integrate a production system with a research and development laboratory, and an administration and policy team to provide a scalable model for Cyberinfrastructure data management evolution and long-term preservation.[40]

While the Foundation did not fund the proposal, it clearly got them, or their governing Board, thinking.

Undeterred, SDSC and UCSDL submitted to NSF and the Library of Congress in a 2005 a DigArch (Digital Archiving) proposal calling for:

> … demonstration of a software technology that will comprise … a preservation environment for a film/video collection that includes other related multi-media content such as audio, transcripts, annotations, related and introductory Web pages, and descriptive, technical, and rights metadata currently being captured in a FileMakerPro™ database … [an environment that would allow one to] take an existing video production workflow … being used for producing and Web-casting video content in a small-scale studio setting, and to integrate it with a digital preservation life-cycle management process that will enable the digital content to be archived for long-term preservation.[41]

[40] "CHRONOPOLIS: Federated Digital Preservation Across Time and Space." Unpublished proposal (May 2004).

[41] "Digital Preservation Lifecycle Management: Building a Demonstration Prototype for the Preservation of Large-Scale Multimedia Collections." Unpublished proposal (March 2005).

The funded project allowed the partners to develop a long-term preservation system that could be inserted into an enterprise-grade television production cycle without disrupting that cycle. As a consequence, in two year's time 300 hours of digital video were archived, along with unedited footage, accompanying audio, transcripts, annotations, related Web pages, and production information.

In 2006, SDSC, UCSDL , and their partners at NCAR and the University of Maryland submitted a retooled version of the Chronopolis proposal to the Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP). Although specifically focused on NDIIPP data collections, this version of Chronopolis was very similar in its vision to that version proposed to NSF two years earlier: namely to provide "… distributed storage, data replication, storage management, and core ingestion, curation, and preservation tools and services"[42] in support of digital data collections stewarded under the aegis of NDIIPP. As noted in the model below (Figure 1), data copies are reposited in three geographically-dispersed physical locations, linked by grid-enabled cyberinfrastructure under varying degrees of access control: bright, dim, and dark. Moreover, any of the three sites can be bright, dim, or dark for a particular data collection. The funded project has two large collections—the social sciences data collection at the Interuniversity Consortium for Social and Political Research (ICPSR; 10 Tb) and the California Digital Library Web-At-Risk collection (25 Tb)—under active management.

[42] "Chronopolis Distributed Storage and Preservation Infrastructure in Support of the NDIIPP Partners Network." Unpublished proposal (August 2006).

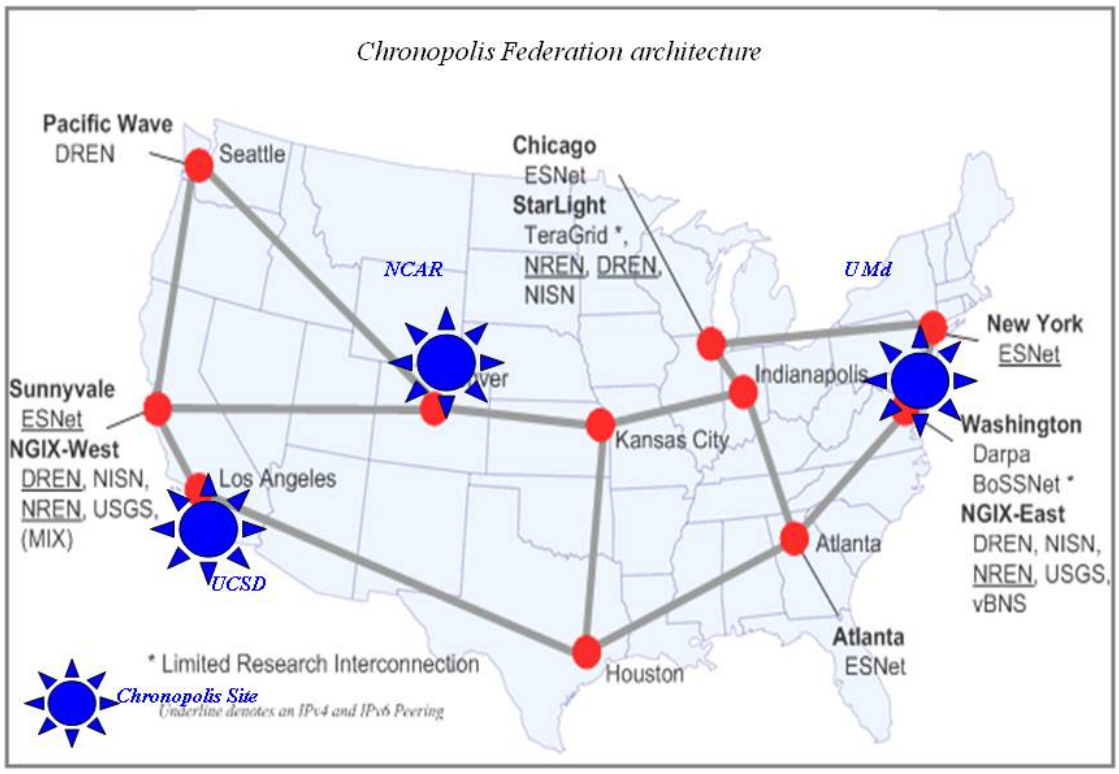

Figure 1[43]

In 2007, the U.S. National Science Foundation issued its "DataNet" rfp, an ambitious call for proposals to create sustainable and extensible long-term preservation services needed to sustain the digital objects (broadly construed) that support science and engineering research and education. In addition, the rfp envisioned this infrastructure being managed and governed by "… new types of organizations … [that] … integrate library and archival sciences, cyberinfrastructure, computer and information sciences, and domain science expertise …"[44] In a rushed, and I believe ultimately divisive process, twenty-seven pre-proposals were submitted, of which seventeen were "qualified"

[43] Fran Berman and Brian E.C. Schottlaender. "'Data' Stewardship." Unpublished presentation (August 2006).
[44] DataNet Program Solicitation NSF 07-601.

for consideration, including a UCSD-led proposal titled "The Data Trust Alliance." Distressingly, ours was not amongst the seven initiatives invited to submit full proposals. Only two of those initiatives will, in turn, be funded in this round. Two or possibly three will be funded in a second round, although one will be required to start the process over from the beginning. UCSD is currently evaluating whether we are interested in doing so.

LOOKING TOWARD THE FUTURE

In 2007, UCSD—under the leadership of its Vice Chancellor for Research, Art Ellis—assembled a campus Working Group to evaluate the future of, and possible future funding models for, SDSC in light of the fact that NSF Core funding for the national supercomputing centers had come to an end. Represented on the group were Administrative Computing, CalIT2, the Libraries, SDSC, and various data-intensive academic domains (Biology, Engineering, Medicine, Oceanography, Pharmacy). The Working Group's terms of reference included:

- determine how the campus should envision the SDSC's role in the near future;
- develop a set of future scenarios and an evaluation of their strengths and weaknesses as potential blueprints for the future of SDSC; and
- be informed by SDSC's connections to other campus units … to the UC system; to the national laboratories; and to the broader national and international user community.[45]

Departing from the observation that "Success in the Information Age can be measured by the precision, power and breadth of available tools and the knowledge of the people who use them,"[46] the Working Group's report notes that SDSC provides the best of both. Of various future scenarios

[45] "San Diego Supercomputer Center Working Group (SDSCWG) Report." Unpublished report (March 2008): p.1.
[46] "San Diego Supercomputer Center Working Group (SDSCWG) Report." Unpublished report (March 2008): p.3.

explored, the Working Group has come out in favor of "integration" at the programmatic level with various other campus agencies operating in the same space, noting that "If … carried out properly, UCSD is poised to become … [a] national model for Research CI, given the collective partnering strengths of ACT, Calit2, SDSC, and the University Libraries."[47]

A separate CI Design Team (CIDT) has now been charged with imagining, assessing, and recommending the best technical architecture for effecting such an integration. The Team comprises representatives from all major computing stakeholders on campus including SDSC, Academic Computing, Administrative Computing, CalIT2, and the UCSD Libraries. The CIDT began its work by surveying campus researchers in order to ascertain their unmet needs in the areas of computing, software, data management, networking, and cyberinfrastructure (CI). A summary of the survey's most significant responses follows.

- Need for Storage Services
    - Backup (88%)
    - Capacity (70%)
- Interest in Co-location (54%) and Condo Clusters (46%)
- Need for CI Services
    - Processing (70%)
    - System Administration (48%)
    - Cluster Administration (38%)

[47] Ibid.

- Need for CI expertise
    - Database and Data Management (62%)
    - Visualization (58%)
    - Portals/User Interfaces (55%)
- Need for Software: Matlab, SPSS, iRODS, IDL, GIS[48]

In thinking about how to address these needs as effectively and efficiently as possible, the CIDT has converged upon the following technical architecture concept (Figure 2). While still evolving, it does convey a flavor of the group's thinking at this point in time.

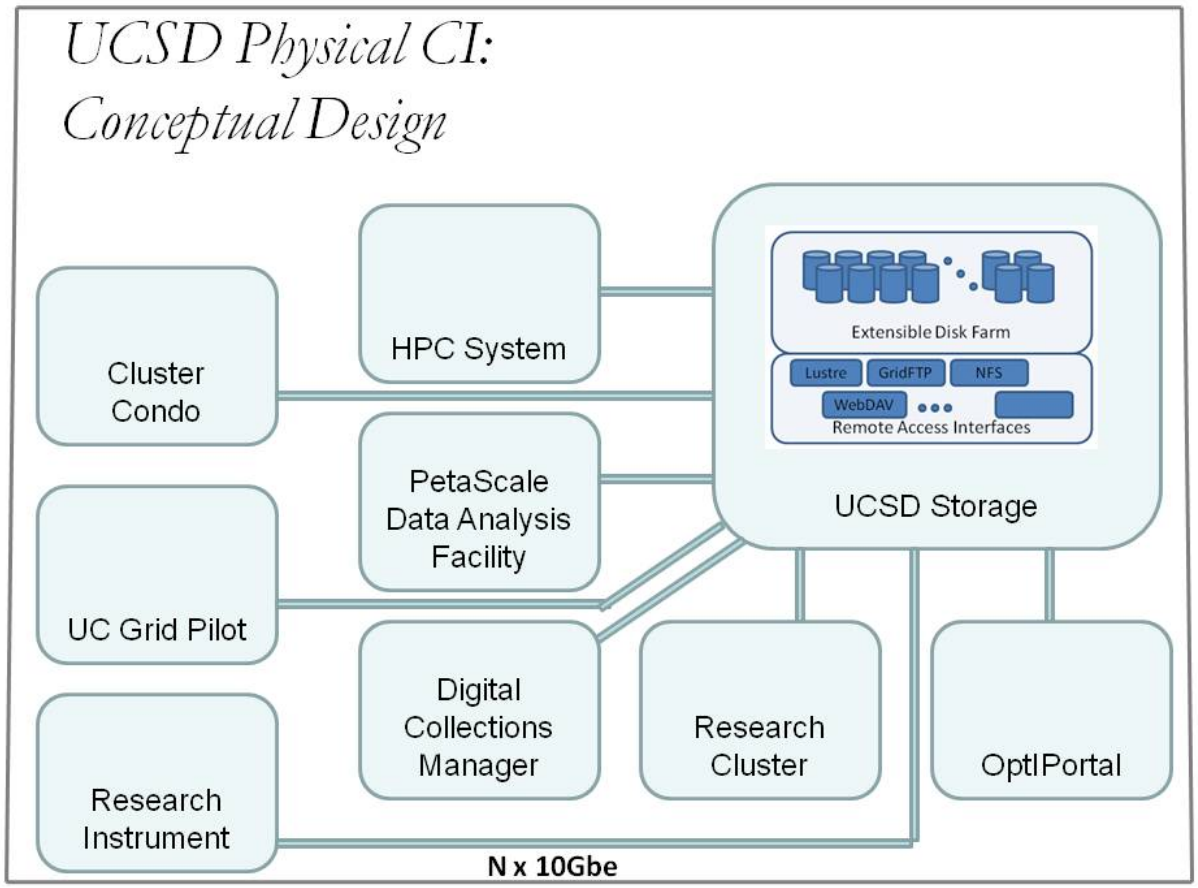

Figure 2[49]

[48] Francine Berman. "Next-Generation SDSC." Unpublished presentation, with slight revision by the author (June 2008).

In parallel with the work of the SDSC Working Group and the CI Design Team, Dr. Francine Berman has been leading her management group through a strategic planning exercise, the intent of which is to imagine and articulate the shape of the "Next-Generation SDSC" (NGSDSC). NGSDSC is designed to accelerate cyberinfrastructure-enabled research and education efforts. Given a prevailing environment that features, on the one hand, "unlimited" data via the Internet, sensors, scientific instruments, and other tools, and, on the other, "unlimited" computation via university clusters, compute clouds, terascale and petascale supercomputers, etc., SDSC is focusing on providing cyberinfrastructure services to empower researchers and educators to do something useful with these riches. Medium-term strategic directions of the Center include:

- Cyberinfrastructure for Emerging Computational Platforms
    - Resources for large-scale data analysis
    - Shared "Condominium" clusters and power-efficient co-location facilities
    - Support for the University of California Grid
- Comprehensive Data Cyberinfrastructure
    - Data services
    - University-level archival repository and storage backbone
    - Integrated data lifecycle management
- Innovation for Next-Generation Cyberinfrastructure Development and Use
    - High Performance Computing on next-generation compute platforms
    - Innovative environments for Cloud platforms
    - Green computing[50]

[49] Phil Papadopoulos and Mike Norman. Unpublished presentation (April 2008).
[50] Francine Berman. Personal communication to author.

With the proper sorts of support, these three focused emphases will combine to produce the "broad impact [and] deep-impact research CI"[51] needed to position SDSC, UCSD, and the University of California as continuing leaders in the development and management of cutting-edge technologies in service to the world's grandest scientific challenges.

"The proper sorts of support," in this context, are not simply financial. On the contrary, as noted by Berman at the outset of this paper, cyberinfrastructure is a "coordinated aggregate." The "body of units … somewhat loosely associated with one another"[52] that constitute UCSD cyberinfrastructure organizationally—SDSC, Academic and Administrative Computing, CalIT2, the Libraries—will need to work together like the components of a finely engineered timepiece if they are not only to succeed, but to amount to more than the sum of their parts.

The next generation of cyberinfrastructure being planned and developed at SDSC and elsewhere is critical for addressing the next generation of E-science challenges. Reducing energy consumption, preventing infectious disease and managing pandemic, predicting earthquakes and tsunamis, mitigating threats to international security: all require the processing, management, and evaluation of massive amounts of data across vast distances and over very long periods of time. In short, all require cyberinfrastructure … and all are uppermost in the minds of today's researchers and universities.

Acknowledgements: The author is grateful to his UCSD colleagues Francine Berman, Luc Declerck, and Mike Norman for their contributions to this paper.

[51] Francine Berman. "Next-Generation SDSC." Unpublished presentation, with slight revision by the author (June 2008).
[52] *Merriam-Webster's Collegiate Dictionary*. 10th ed. (2002): p. 23.