The Responsible Stewardship of Research Data: A Roadmap for the University of California, Merced

September 4, 2012

Emily Lin
Head, Digital Assets
UC Merced Library

This paper frames the case for a strategy to address the responsible management, curation, and stewardship of the research output of the University. However, it is the author's observation that as with research data, individuals and units across this institution have been left largely to their own devices to manage the data they produce or use. As a result, this institution loses productivity and efficient use of resources due to siloed approaches to data management; is at risk for not meeting security and compliance needs; and is unable fully to leverage data for effective assessment and decision-making.

The Library is concerned with the responsible stewardship of research data in order to fulfill its core mission of supporting the current and future needs of scholars who require access to information. Nonetheless, a strategy should be considered in the context of the larger need for responsible and effective data management institution-wide, which is critical for the long-term success of the University.[1]

I. Background and Drivers

In September 2005, as UC Merced opened its doors to its first class of students, the National Science Board (NSB) issued a report, "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century" (http://www.nsf.gov/pubs/2005/**nsb**0540/), that drew attention to the

[1] UCSD's Research Cyberinfrastructure (RCI) Design Team also recognized that their proposed CI design would benefit the teaching and business components of the university, but chose to focus on research data, noting that "scaling to the entire campus community would add considerable expense and additional design time" and "would need to consider replacement of existing infrastructures" ("Blueprint for the Digital University," April 2009, http://research.ucsd.edu/documents/rcidt/RCIDTReportFinal2009.pdf, p. 29). Given UC Merced's nascent infrastructure, the considerations may be different, although priorities and directions still need to be well defined.

significant investments the National Science Foundation (NSF) makes in the creation and maintenance of digital data collections and the need for a corresponding technical, financial, and policy strategy to ensure that maximum benefit of those investments will be realized over time.

One outcome that grew out of the National Science Board's recommendations was the institution of the requirement in January 2011 that all research proposals submitted to the NSF should include a data management plan. The purpose of the requirement was three-fold:

- to include within the peer-review process evaluation of how activities that will generate digital data will "meet the standards, norms, and expectations of the community";
- to determine if the proposed budget adequately supports the data management plan; and
- to enable the NSF to track the PI's effectiveness in implementing the data management plan.

Attention to data management and sustainability has not been isolated to the National Science Foundation. Shortly after the NSF implemented its requirement, the National Endowment for the Humanities (NEH) announced that applicants to its Digital Humanities Implementation Grants must meet a similar requirement. Back in 2003, the National Institutes of Health established the expectation that research data from NIH-supported studies would be released and made accessible for use by other researchers; proposals seeking $500,000 or more would be required to include a data sharing plan.

This past March, the White House announced its "Big Data Research and Development Initiative," drawing together commitments across six Federal agencies to advance the ability to "collect, store, preserve, manage, analyze, and share huge quantities of data" in order to accelerate scientific discovery and transform teaching and learning.

Data has become a type of currency underpinning operations in business and social enterprises in the twenty-first century, as well as within the research and academic enterprise. Good data is recognized as a valuable asset that enables or even drives learning, decision-making, and productivity. If UC Merced is to establish itself as a twenty-first century research university, it, too, must meet the obligations and opportunities of ensuring that the significant investments made in the generation of research data at the University will reap maximum benefit to researchers and scholars over time. What is at stake:

- **Funding.**
  Since its founding, UC Merced has received nearly $22 million in NSF awards, which comprise 26% of the total extramural funding awarded to UC Merced. Funding from other federal agencies, including the NIH, have amounted to an additional 25% of UC Merced's total extramural awards. As effective management practices for long-lived data and data sharing become criteria for garnering and maintaining research funding, it is urgent that the University also has the technical and financial strategy as well as policy in place to meet those expectations. Since granting agencies are also funding data management costs, it is in the interest of the University to establish strategies that will make the best case for and best use of extramural funds.

- **Credibility.**
  Great universities have been known for the research collections they own. In a digital data-driven world, great universities will be known for the research collections they have produced and to which they have provided access. Questions of ownership aside, assigning the institutional name of UC Merced as the source of datasets or data collections enhances the reputation of the University as those data are shared, accessed, and reused by others.

  Ensuring that data is available for audit or reuse is also critical to protecting the credibility of research findings and the reputation of the research institution. Recent high profile cases of falsified data or irreproducible research claims have underscored the need for institutions to take steps to reinforce responsible conduct of research with requirements for proper management, archiving, and dissemination of research data.[2]

- **Recruitment and Retention.**
  Reputation shapes the ability of the University to attract talented researchers—both students and faculty. Providing the infrastructure and services that support research activities is critical to the retention of researchers. Baseline needs, as identified by the survey results of

---

[2] Goozner, Merrill, "Duke Scandal Highlights Need for Genomics Research Criteria, "*JNCI Journal of the National Cancer Institute* 103, no. 12 (2011): 916-917, doi: 10.1093/jnci/djr231; David Dobbs, "Marc Hauser News: A Settling, or Pre-Quake Tremors?" *Wired Science Blogs/Neuron Culture*, May 2, 2011, http://www.wired.com/wiredscience/2011/05/marc-hauser-news-a-settling-or-pre-quake-tremors/; Benedict Carey, "Fraud Case Seen as a Red Flag for Psychology Research" *New York Times*, November 2, 2011, http://www.nytimes.com/2011/11/03/health/research/noted-dutch-psychologist-stapel-accused-of-research-fraud.html; Ed Yong, "The Data Detective," *Nature* 487 (July 5, 2012): 18-19, doi:10.1038/487018a.

academics at this institution (see "III. Status Quo at UC Merced" below), have not been well addressed. Many institutions are still in the formative stages of developing the "cyberinfrastructure," including digital curation services and policies, to support "e-research." Developing and sustaining a leadership edge in this area will translate into a competitive advantage for the University.

- **Impact.**
  Usefulness of a dataset may well extend beyond the life or scope of an individual research project. As cited in the "Long-Lived Digital Data Collections" report, the Protein Data Bank is an example of a small research data collection that has evolved into a premier reference collection supported and utilized by researchers around the world. Data that has been costly to generate or collect, such as image or instrument data, may have high value to other researchers. Synthesis science, correlational studies, and other types of research may aggregate and build upon data generated by a variety of sources or over time.

  The University has an obligation to provide broad access to research products that have been publicly funded. The NIH and the NSF both have data sharing policies that outline the expectation of the timely release of data created or gathered in the course of sponsored research. By preserving and making accessible research data, the University also has the opportunity to inform the broader public, to enrich and foster further discovery, and make significant, immediate, as well as long-term impact.

II. Initiatives at Other Universities

UC Berkeley Information Services & Technology launched Research Hub (https://hub.berkeley.edu) in September 2011, a web-based service that allows anyone in the campus community to store and manage files (up to 10 GB free) and create sites for collaboration among research teams and projects. The service quickly exceeded its first-year goals in terms of user adoption: there are currently 2200 users of the service, with a high percentage of academic departments utilizing the service for either administrative or academic purposes. The Hub runs on a commercial, open-source platform backed by central storage and backup systems and has thus far shown to be very cost-effective. The objective behind the Hub is to support the full lifecycle of digital content and its use.

UC San Diego issued a "Blueprint for the Digital University" in April 2009 proposing the design for a campus-wide research cyberinfrastructure (RCI). A 2010 Cyberinfrastructure Planning and Operations Committee report develops a business plan to implement five RCI elements: a collocation facility; centralized

data storage; data curation; condo clusters; and a research network. An RCI Oversight Committee currently oversees development and sets policies, while an Implementation Team comprised of members of the UCSD Libraries, San Diego Supercomputer Center, CalIT2, and Administrative Computing and Telecommunications (ACT) are managing five pilot projects as part of a Research Curation and Data Management Pilot program begun in 2011. The campus is fully funding the five pilot projects but will need to develop a sustainable funding model. See http://rci.ucsd.edu for further details.

Purdue University convened a campus-wide group chaired by the Dean of Libraries and the Vice Presidents of Information and Research in 2010 to examine the data management needs and practices of its researchers. Based on the report of the committee and results from faculty meetings, Purdue is creating an institutional data repository service using Purdue's locally created, open-source HUBzero software. The Purdue University Research Repository (PURR) is a joint effort of the Libraries, IT, and Office of the VP for Research, with HUBzero costs divided among the three partners for three years. See Michael Witt, "Co-designing, Co-developing, and Co-implementing an Institutional Data Repository Service," http://dx.doi.org//10.1080/01930826.2012.655607.

Cornell University proposed a model in October 2010 for a Research Data Management Service Group that will "present a coherent set of services to researchers" regarding data management planning, services available on campus, and a single point of contact for specialized assistance. The group is jointly sponsored by the Senior Vice Provost for Research and the University Librarian, and has a faculty advisory board and a management council with members from the University Library, Center for Advanced Computing, CISER, Weill Medical College, and CIT. See http://data.research.cornell.edu for more information.

The University of North Carolina at Chapel-Hill's Provost charged a campus-wide Task Force on the Stewardship of Digital Research Data in January 2011, chaired by the Dean of the School of Information and Library Science. After conducting an environmental scan of data stewardship policies and trends as well as surveying the campus, the Provost's Task Force issued a report in February 2012 that outlines a set of principles and actions for the campus to consider. See http://sils.unc.edu/sites/default/files/general/research/UNC_Research_Data_Stewardship_Report.pdf for the report.

At a national level, research universities are forming a federation called the Digital Preservation Network (DPN, http://d-p-n.org) to create a distributed, networked approach to preserving the complete scholarly record by replicating nodes at different institutions. The DPN was launched in Spring 2012 and is

championed by James Hilton, UVa Vice President and CIO. Over fifty institutions have committed initial support (a $20,000 initial commitment) and members from the UC Libraries' leadership have been involved in the development of the initiative. UC Merced will need to continue to stay apprised of developments within UC and at the national scale and consider strategic involvement in these initiatives.

III. Status Quo at UC Merced

After conferring with the Vice Chancellor for Research in fall 2010, the UC Merced Library worked with the Office of Institutional Planning and Analysis to conduct a campus-wide survey of researchers regarding their data management needs. The survey was issued in January 2011 to 296 people within the academic series ranging from postdoctoral researchers to full professors. A total of 72 people responded (a response rate of 24%). Out of 144 ladder-rank faculty (assistant professor to full professor positions, as of August 2011), 46 responded (32%). Participants were asked to rate 16 different types of data management needs on a five-point scale in terms of importance (very important - not important).

Among the sixteen needs listed on the survey, the three that were identified by the largest number of respondents as "very important" are "Reliable, redundant short-term storage of data (1-5 years)" (65% of respondents); "Ensuring authenticity and integrity of the data I produce" (47%) and "Allowing or controlling access to data" (46%). Seventy-four percent of respondents indicated that long-term preservation of research data was very important or important.

Top 10 Data Management Needs Ranked by UC Merced Researchers

1. Reliable, redundant short-term storage of data (1-5 years)
2. Long-term preservation of research data (beyond 5 years)
3. Transferring research data to storage
4. Transferring research data from storage to desktop/cluster
5. Allowing or controlling access to data
6. Ensuring authenticity and integrity of data
7. Maintaining data/format compatibility
8. Sharing data with colleagues
9. Accessing data from national or community repositories
10. Publicizing or enabling discovery of my data

Only a slight majority (52.8%) of respondents answered affirmatively that their current needs for research data management are being met. Among the 19 comments submitted regarding what needs were not being met, many commented that they were handling data management needs on their own, using

personal machines or even Dropbox for storage. Several expressed a desire for managed, centralized data backup service. A related need expressed in comments was the ability to share data with external groups, such as research collaborators, reliably and conveniently. Finally, a few comments called explicitly for the need for consultation and expertise on data management.

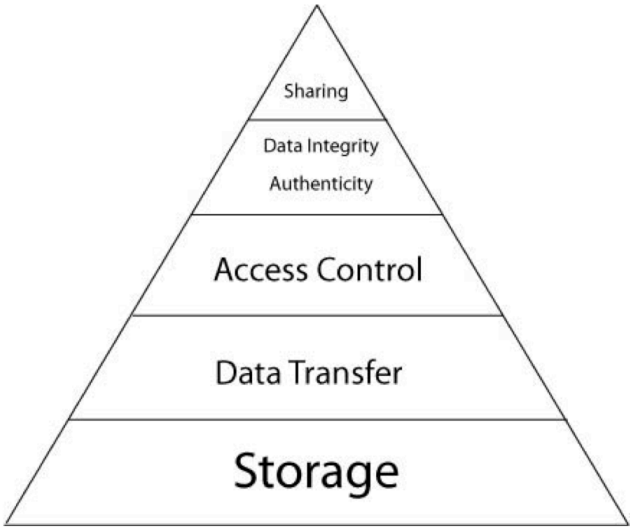

Figure 1 A "Maslow's Hierarchy" of Data Management Needs

At the time of the survey, only one-third (33.3%) of respondents were aware of obligations being imposed by the NSF and the NIH in terms of data management and access. Less than 20 percent specify a budget or research allocation for data management for research projects. As one respondent commented, "We specify one but it is almost always far too small to account for the costs of curation and data access." Others only budgeted "for storage hardware, but not for maintenance or backup" or assumed that "such things should be considered part of infrastructure (indirect cost)."

UC Merced researchers are producing data across the spectrum of data types; while the majority produce electronic text and numerical data, a significant number (nearly 70% combined) are producing images, video, and audio as part of their research—all of which have more complex requirements in terms of management and storage. In addition, 47 percent of respondents indicated they are generating or maintaining research data in non-digital formats (paper, photographs, video or audio tapes, slides, etc.).

Of significant concern when considering risk management and reliable data storage practices, nearly 60 percent of respondents host their research data on local equipment. Very few indicated that they host research data on any networked storage (26% indicated "servers maintained by the University) or on storage residing geographically outside of the University (17% indicated hosted

solutions). The majority of respondents (39%) indicated that they currently share or plan to share data using e-mail.

Funders emphasize access to and availability of data as one of the drivers for their data management policies. The National Science Board issued a "Memo on Digital Research Data Sharing and Management" (December 14, 2011) that stated its commitment

> to the development, implementation, and assessment of policies that promote efficient management of, and broad access to, digital research data that result from NSF-funded activities. This commitment includes sharing of results, data, physical collections, and other supporting materials created or gathered in the course of NSF-funded research.

However, the survey indicates many UCM researchers place lower priority on the needs related to publicizing or disseminating their data. For example, respondents ranked creating metadata, which is essential for identification and discovery of data, as well as creating persistent identifiers for citation, lowest in terms of importance, perhaps because the purpose and methods for those activities are not well understood.

It is clear that the campus needs to provide reliable solutions for both short- and long-term data storage and management. The campus also requires solutions that will facilitate the secure transfer, sharing, and control of research data. Researchers must be well advised of their obligations, of how to plan for their needs, and of the services and resources available to them.

IV. The Role of the Library

The purpose of a university library is to ensure that the researchers it serves can access the information they need. Traditionally, that purpose has been translated into two core missions: to collect and preserve information resources, and to provide access to those information resources. The transition to digital information, digital access, and digital modes of research has transformed how and what libraries collect, preserve, and deliver. An information resource no longer needs to be held locally for users to access it, although keeping multiple copies in distributed locations remains the best method for preserving it. Digital modes of research and production have altered radically what kinds of information researchers now use and what they do with it: from generation or discovery and collection; to processing and analysis; to interpretation and communication of new knowledge.

Because of this transformation, the library is no longer omnipresent and essential at the beginning stages of this lifecycle. However, because the library still holds responsibility for ensuring access to information for the user—as the neutral, honest broker who acts in the interests of the institution to secure and deliver resources in the most efficient and effective way—the library is acutely concerned with how practices and policies at the beginning stages affect the overall lifecycle, from creation to transmission, of information and knowledge.

Through managing the digitization of a complex collection of mixed format information resources and subsequent digitization projects (encompassing text, image, artifacts, audio, and video), the UC Merced Library has experience and expertise in identifying and adopting standards for data and metadata creation; in managing the secure storage, backup, and transfer of multi-GB data files; in enabling access to and discovery of digital data collections through aggregators; and in navigating and establishing terms of ownership, access, use, and sharing. In addition, the Library has conducted limited pilot projects supporting the creation and publication of collaborative digital research products by UC Merced faculty. Given this expertise, the Library is well positioned to advise and support researchers in terms of data management planning and implementation. Given its responsibility to ensure that the record of UC Merced research will endure and remain accessible to other researchers, the Library has a vital interest in doing so.

V. Actions Required and Questions to Be Answered

A. Establish and communicate clear policies for data ownership, retention, and sharing. Require every research project to have a data management plan and determine a means for tracking compliance.

Buried in the UC Academic Personnel Manual (APM-020), Regulation No. 4, Sec. 2.5 states

> All such research [for the benefit of Federal, State, industrial or other projects] shall be conducted so as to be as generally useful as possible. To this end, the right of publication is reserved by the University... A report detailing the essential data and presenting the final results must be filed with the University. Notebooks and other original records of the research are the property of the University.

At present, neither the UC Office of the President Office of Research and Graduate Studies nor the UC Merced Office of Research websites, however, explicitly state this policy, nor do they provide policies or guidance on how research results should "be filed with the University." While a UC retention policy exists for specific "Administrative Records Relating to Research," there should be guidance or policy on

- how long research records in general should be retained;

- expectations surrounding record keeping practices such as data security;
- how retention applies to written correspondence such as mail and email;
- requirements for access; and
- transfer should a researcher leave the University.

Specific examples of other institutional policies include the Ohio State Research Data Policy (http://orc.osu.edu/files/2011/01/ResearchDataPolicy.pdf) and the Harvard Retention of Research Data and Materials Policy (http://osp.fad.harvard.edu/content/retention-of-research-data-and-materials).

If the University owns the research data, the University needs to take responsibility for the data, especially for datasets that have no disciplinary or community "home." If researchers claim in their funding proposals that the University will maintain or manage their data products, a formal deposit agreement should be in place and responsibility assigned from the outset of the project to oversee the fulfillment of the agreement. Ultimately, because the University bears responsibility, requiring a data management plan for every project will ensure that expectations and outcomes are clearly defined for all parties.

Question(s):

- What data is retained and for how long?
- When is data publicly accessible?
- What happens when a researcher leaves the University?
- What provisions will the University make to enforce policies on data? Who will track compliance?

If no action is taken: External funders have data management and data access requirements, and may ultimately audit compliance. The University risks a high-profile case such as what occurred at Harvard, where research findings were subject to question (and ultimately research misconduct was determined to have occurred) to prompt inquiry into and audits of research data policies and provisions. Increasing attention to and demand for the data supporting research findings heightens the urgency for clear communication of guidelines and expectations.

While UC Merced is young, there are already researchers who have retired or moved on to other institutions. What is there to show for the considerable investment this institution has made in bringing researchers to the campus and supporting their time at the institution?

B. Establish the capacity and expertise to support the full lifecycle of research data generated at UC Merced

In April 2010, a faculty taskforce submitted a proposal for addressing the research computing needs of the campus with the strong statement that current "UCM-IT has consistently declined to place research computing as a priority service." In addition to identifying specific needs, the proposal outlined a vision for 1) aggregation of research computing resources to achieve a higher overall level of support and economy of scale; 2) staff and resources directly accountable to the users and 3) flexibility to support diversity of needs and different revenue sources.

Since no action was taken on the first proposal, in fall 2011 the Graduate & Research Council of the Academic Senate formed a taskforce, formally charged by the Vice Chancellor of Research, to "reconsider this topic in a formal way." Chaired by the dean of the School of Natural Sciences and composed of one faculty representative from each of the three Schools as well as the author of this report, the small taskforce solicited input from faculty to identify primary "research drivers" and research computing needs.

The findings correlate with the responses to the 2011 survey of data management needs and with the needs identified by the RCI design team at UCSD. In other words, the research computing needs of the UC Merced faculty are not unique. Yet if UC Merced is to achieve status as a top-tier research university, it cannot afford further inaction in addressing these fundamental needs. One faculty member's comment astutely summarizes the needs expressed by the whole:

> There is no dedicated full-time versatile staff support to maintain, upgrade and administer computing systems, no data back-up facility, and no currently functioning shared UNIX-based computing facilities. Faculty have to directly manage staff and system maintenance, and outsource expensive administration services. Like core facilities on other campuses, these facilities should really run themselves. Full-time staff salaries must be high enough to recruit expert talent to Merced.

The needs that emerge from the aggregated feedback from faculty (see Appendix 1) are:

1. Shared or centralized computing resources, possibly outsourced
2. Reliable, backed-up data storage
3. Shared UNIX/Linux-based high-performance computing clusters
4. High-performance, high-speed networking
5. Dedicated expertise with UNIX/Linux knowledge not only for active maintenance of systems, but also to consult and advise on optimal utilization of resources, as well as data modeling and analysis
6. Data curation resources to support management, description, and sharing of data

While there is consensus in support of shared or centralized computing resources, there is mixed opinion on the extent to which services can be outsourced. Outsourcing to UC (e.g. LBNL, SDSC) or to commercial services for storage and high-performance computing services may be the most viable and cost-effective solution, although questions regarding security and service levels will need to be addressed. However, the campus will still need to make significant local investment, in particular to ensure reliable, secure, and high-speed transfer over networks and to provide dedicated computing expertise and institution-specific data curation. At present, the University provides no "central repository" to serve up data that needs to be made accessible, for which there is no other convenient external home.

Question(s):

- What is the vision for information technology management for the campus?
- What services can or should be outsourced, and what is the University able to invest in local capacity?
- What current IT resources on campus can be leveraged to support research computing and data curation needs? What gaps in resources and expertise will need to be filled?
- What can be done in the immediate term before a vision for IT can be developed and implemented to address the pressing needs of researchers who work within grant cycles and must meet external requirements?
- How will the University ensure accountability and responsiveness of service providers to users, in particular given diversity of needs and potential disparities between high-demand and low-demand users?
- What will be the baseline expectations and levels of service to be met by service providers?

If no action is taken: Researchers will continue to find their own solutions to their research computing and data management needs. The situation is untenable: hard drives and servers will corrupt or fail without monitoring, redundancy, and refreshment; valuable data representing time and effort and which may be unique and irreproducible will be lost. Valuable time, including time spent by students and research assistants, will continue to be wasted on IT deployment and maintenance instead of being spent tackling research questions. There will be continued waste in equipment and energy costs: according to the UCSD RCID report,

> individual clusters often run at an average utilization of 3-10%... centralized infrastructure can be operated more economically, both from a human perspective and from an energy and cooling perspective. Operating machines in a facility designed to support computing and

storage can incur operating expenses that are a factor of 2-10 less than ad hoc deployments.[3]

Finally, without an overall strategy and implementation for the campus, expertise and resources will continue to be unevenly developed and distributed. The Digital Assets unit of the Library has been in contact with UC Berkeley and UCSD regarding their cyberinfrastructure initiatives and both are willing to provide some extent of services to UC Merced. The Library is identifying other third-party services that can be communicated to researchers, but this is still a stopgap approach.

C. Establish a structure with clear authorizations, roles, and sustainable funding for developing, communicating, and delivering services in support of research, including data curation, to users.

A single point of contact, such as UCSD's RCI group or Cornell's Research Data Management Service Group, and consistent messaging about policies and services needs to be established with clear support from the Provost on down. Modes of communication to principal investigators need to be defined and optimized. Constituents include the Office of Research, the Schools, the Library, and IT. Roles for each need to be defined, and their contributions to total cost, as well as the costs individual researchers are expected to bear, also need to be defined.

Defining roles and handoffs in the process from inception and development to completion and closeout of a research project is critical to the successful management of research activities and products. Once the roles and processes are well defined, the systems and necessary integration across entities can then be appropriately developed and supported. With the transition to the digital age, universities are tapping data applications not only to enable research administrators to more effectively manage institutional activities, but also to enable researchers to build upon or develop new projects or collaborations based upon existing research activities and expertise.

The University, with proper planning, can make more effective use of data for both internal and external purposes. The Office of Research is currently implementing an electronic research administration tool to develop and track proposals and grants and support strategic decision making for research activities. Such a tool could be used to track components such as data management plans and data agreements, but this has not yet been addressed. Many universities have also implemented tools like VIVO (http://vivoweb.org/) to support and enable discovery of an institution's researchers and researcher output—addressing another piece, or the tail end, of the research lifecycle.

---

[3] "Blueprint for the Digital University," http://research.ucsd.edu/documents/rcidt/RCIDTReportFinal2009.pdf, p. 31.

Institutions across the globe[4] are utilizing such a platform to aggregate data about their researchers and research activity to promote internal and external uses such a new project generation and cross-disciplinary collaborations. The Office of Research should lead the selection and deployment of such a tool, but the Schools and ORUs, business and financial administration, the Library, and IT are also stakeholders with pieces to contribute: data, expertise, use cases, and resources.

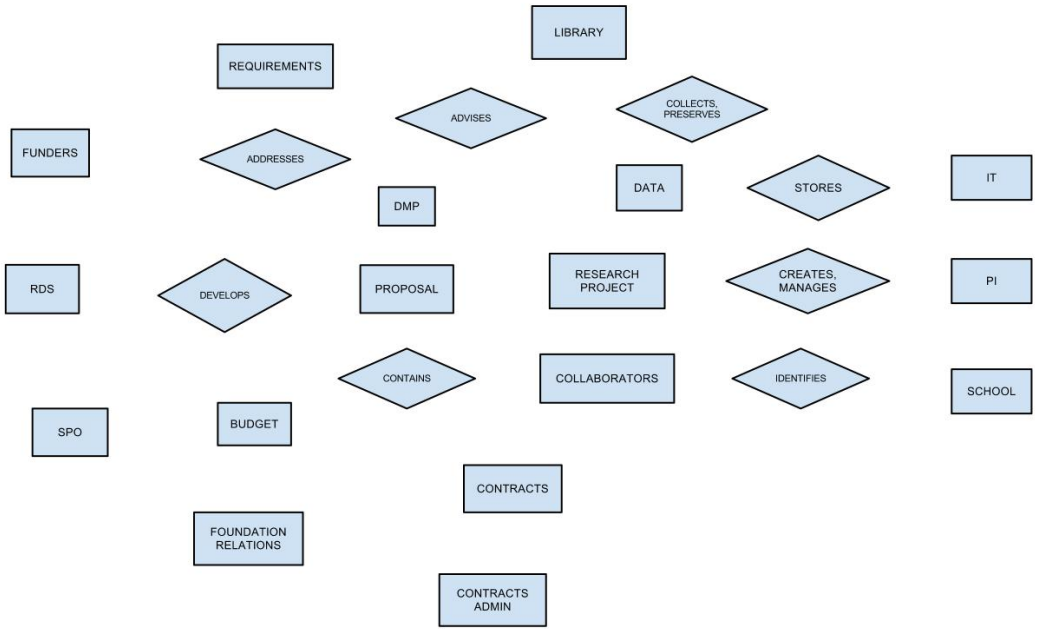

Figure 2 The makings of an entity-relationship diagram for research

Question(s):

- What models for sustainable funding, including tapping external sources, can be pursued?
- How will consensus on a structure, including roles, authority, and responsibilities be achieved? How will such a structure be implemented and in what timeframe?
- What integrating devices, including formal lines of communication and reporting, should exist to ensure overall institutional goals are met as well as the health and success of all members of the institution?

---

[4] While VIVO was developed at Cornell, adoption and implementation has occurred around the world. Notable recent implementations include the University of Melbourne (http://www.findanexpert.unimelb.edu.au) and Eindhoven University of Technology (http://vivo.libr.tue.nl/). In addition, library/information science professionals are utilizing the tool to aggregate data and further discovery: see the International Researcher Network Visualization (http://nrn.cns.iu.edu).

If no action is taken:

UC Merced is at a critical juncture in its development as a research university. Without action to address these issues, the long-term success of its endeavors is at risk, and opportunities will be lost. With the right focus and the right pieces in place, the institution will be able to establish a culture that runs on a solid foundation of good data practices and responsible data stewardship to garner continued success and deliver the full impact of its contributions to the world.

Appendix 1. Compilation of excerpted, anonymized feedback solicited and gathered by D. Ardell (SNS) and F. Rusu (SoE) from faculty on research computing needs (Spring, 2012) and comments submitted by respondents to 2011 Data Management survey

A. Model that will be more cost-effective and keep pace with advancements in technology: shared or centralized; outsourced

- current computing resources on campus are non-existent
- antiquated technology
- a centralized computing model would save me from buying expensive, single use… lab equipment
- the cluster we spent a lot of money on in 2008 is fast requiring replacement parts, etc. and never had a proper back up system
- no real need for the headaches of managing a computing cluster myself
- There is no research computing support presently at UC Merced, in which model faculty can budget and pay recharge to
- Since UCM is a CITRIS-campus, it would be ideal to pursue the creation of more UCM-accounts in other centers such as NERSC, as LBNL is also affiliated to CITRIS. This will allow for additional platform options and specialized resources for faculty at minimal or no cost.
- Access to shared computing resources
- Support via staff system expertise and flexibility with allocation of computing resources
- The only real paths to success are going to have to be very non-traditional, like out-sourcing CPU-, memory- and storage-intensive services entirely to google or apple cloud servers owned, managed by them and located elsewhere. Doing so would involve a loss of control and convenience, but we could more easily ramp up and ramp down without building in a big financial commitment.

B. Reliable data storage

- The major item I need is reliable data backup… nightly network backup system
- high-reliability mass storage
- long-term data storage abilities
- server farm with pretty big data storage
- large datasets
- replacement RAID storage array, a reliable automated back up system
- Data backup (redundancy) is a challenge for me at the moment.
- right now I just store data on portable hard drives and disks. We are trying to move to store more data on free sites like Google so that we have

additional off campus backup, but it is impractical for some files.
- I really could use much server space to store, back-up data.
- I would like to see a managed centralized data backup service for use by faculty for reliable storage and backup of data.

C. Shared UNIX/Linux-based high-performance computing clusters

- my lab would find it advantageous to have access to... a cluster platform
- need access to proportionately more cpu cycles. I prefer to have on-campus computers rather than teragrid or similar, because my calculations do not benefit from parallel architecture and I tend to run large numbers of serial jobs at any given time… like to have local machines to avoid long waits in a queue
- interested faculty could buy in to a shared campus cluster by purchasing nodes
- access to a shared cluster for the analysis
- analyses are computationally intensive
- computing research is CPU and storage intensive
- require high-performance scientific computing resources that are currently unavailable at UC Merced… a cluster with 256 nodes will be highly beneficial
- The larger the size of the cluster we can access, the more scientifically/technologically important the problems we can solve
- Computer clusters of 20 nodes and up.
- 60 dual Intel processor, quad core nodes. Each of the nodes has 8 cores, 160 GB of local disk, 16 GB of memory, and an InfiniBand interconnect.
- A computer cluster consisting of 10-20 CPUs with 16 GB each
- Really need to build Beowulf clusters on campus!
- High-performance computing (e.g. > 300 CPUs) which is currently performed at UCSD

D. High-performance, high-speed networking

- Fast I/O to disk, large amount of RAM per node, and high-speed networking will be just as important as the more traditional focus on CPU clock speed and number of nodes
- Access to the file system using NFS, not sshfs (too slow and drops occasionally). ***This is fundamental***

E. Dedicated expertise with UNIX/Linux knowledge not only for active maintenance of systems, but available to consult and advise on optimal utilization of resources, as well as data modeling and analysis

- Dedicated, full-time UNIX research computing staff support for active maintenance and upgrade of systems as well as consultation to help researchers utilize core computing facilities and service. Staff would not

only deploy and maintain local systems but advise on utilization of cloud computing.Da

- Consistent method to fund IT staff who could oversee the health and maintenance of research computers
- I pay for contracted service for bigger job items that are outside of my expertise
- zero infrastructure for maintaining UNIX computers
- need for significant computer support for data reduction and modeling
- time to audit the systems for intruders, update the OS, manage user accounts, etc.
- at least one dedicated staff member for research computing support (system administration, hardware replacement, etc
- dedicated full-time versatile staff support to maintain, upgrade and administer computing systems…Faculty have to directly manage staff and system maintenance, and outsource expensive administration services. Like core facilities on other campuses, these facilities should really run themselves. Full-time staff salaries must be high enough to recruit expert talent to Merced
- Having dedicated staff for each school or type of simulation may contribute to a more "personalized" and efficient assistance.
- "Advanced Service Providers" in the form of access to highly competent system administrative support with flexible competence in supporting a range of platforms and software, but specifically Linux software. In particular, the kind of system administrative support needed is one requiring ongoing professional education to keep up with evolving standards, protocols, and technology.
- Highly specialized support, on a limited time basis (short period of time)
- GIS/Data analysis assistance needs to be an expert level and needs to support the data analysis software faculty use (i.e. Stata) to be useful
- The real bottleneck now is human capital... for example, I am not keeping up with the latest on how to get the most out of the resources I already have: the hardware is improving faster than I am. I expect that I could do a much better job of designing my modeling, spreading calculations across CPU's on individual or multiple machines, etc. Probably the best thing that the campus could do would be to provide training for PI's and key lab personnel to build up our human capital, so we can do things ourselves that leverage the incredible resources we already have available thanks to the rapid improvement in commercially available computing power.
- A system administrator to take care of it.

F. Data curation resources to support management, description, sharing

- Need library to supply data management/software consultations from someone with substantial expertise. Especially for GIS data that many use on campus.
- work generates a lot of data, and really tests our ability to manage data, versions, models, etc.
- for every scenario, grid cell, month, etc., we have a 1000 or more simulations. It adds up fast... just doing 3 scenarios over greater Yellowstone for 150 years involved over 5 billion random simulations.
- We are creating a digital library and working with the NSF CZO group to develop a protocol for data storage that will be used by other universities. That does not address my needs for document management and databases that we use for other projects.
- We are in the process of linking multi-year data from 3 campuses/5 faculty/6 geographical sites and are exploring how best to manage the data and could use help.
- My primary problem will be the curation of complex models of spatial data. These are now in proprietary formats, and find a solution for their long-term preservation and access to them is proving problematic.
- One area of concern even now… is in conveniently sharing data with colleagues outside UCM.
- I have data sets that need to be accessible to the other scholars
- One of my NSF grants require that I make my research data publicly available.
- it is not enough if all components of metadata meeting standards is to be completed by scratch for our research group
- Last, with NSF's new data management requirement, I'm in a bind. I generate a lot of raw output and don't have a straightforward method to serve it. There is no central repository for University based… output that is not associated with community-scale efforts.