UNIVERSITY OF SOUTHERN CALIFORNIA
UNIVERSITY RESEARCH COMMITTEE 2014-2016

Academic Computing - December 7, 2015

This report summarizes the findings and recommendations of the University Research Committee (URC) on the topic of Academic Computing at USC and in particular the overarching question of "how academic computing at USC needs to evolve to take advantage of, and adapt to, technologies becoming available to support communication, computation, data storage, and data sharing." The URC's focus areas for this report were:

- Identifying processes for determining which software are most desired and most used by faculty in their research, representing strong candidates for university level site licenses.
- Determining needs for the effective use of software, including staff support, training, a centralized and simple database of such resource, building user communities.
- Understanding the needs of faculty in regards to data storage and identifying mechanisms for providing data storage solutions as it pertains to research.

The URC consulted several information technology (IT) experts across campus that provided valuable information and insight into USC software, computing, and IT resources. A university-wide faculty survey was also conducted in which faculty were queried regarding their software use and needs. In addition, a University and USC Hospital Operations Internal Audit Report was recently generated regarding data backup and disaster recovery and was considered by the URC. In general, the current URC study focused more on software resources and data storage than computing resources and infrastructure.

The following individuals met with the URC, during which they provided an overview of resources at their respective units and engaged in discussions with the Committee regarding strengths, weaknesses, and plans related to their programs:

- Douglas Shook, PhD [Chief Information Officer and USC Vice Provost for Information Technology Services (ITS)]
- Candace Borland (Director of ITS web services)
- and Joe Cevetello (ITS Assistant CIO)
- Sam Gustman (Associate Dean at USC Libraries, Chief Technology Officer of the Shoah Foundation)

URC Findings regarding Academic Computing

Finding 1 (Information Technology Services):
ITS provides software support across campus units for several common computing and statistics packages such as MATLAB, Mathematica, Qualtrics, and SAS. While the ITS website has information on all supported software, review of the website revealed difficulty in locating this information in many cases. Some of the available packages, notably SAS, are difficult to install, but currently there is no standard user help or training for any of the "supported" software. Support *is* available at the ITS UPC location for walk-in users.

Finding 2 (ITS Funding):
Funding for ITS-provided software is in part contributed by various Schools and Colleges across campus and by ITS discretionary funds. The ITS procedures for acquiring software site licenses vary, and is based on individual (or groups of) faculty putting requests to their Deans in a process that is not well advertised to faculty members. ITS does not have a more formal model for determining which software packages are needed and used for USC research.

Finding 3 (ITS purchase of Adobe Creative Cloud):
ITS had been engaged in discussions within campus and with other universities to investigate methods for acquiring Adobe Creative Cloud site license. This particular package is mentioned because it has very high demand across USC, but is extremely expensive. Adobe site license costs are very high because Adobe charges per user, not per "seat". Therefore, even though the cost per user is one-tenth of retail cost, the total is unaffordable by this model. ITS had made the decision to instead purchase a group of individual license ($32,000) instead of an institutional license ($27/user, ~$900,000).

Finding 4 (Bioinformatics Program):
The Bioinformatics Program provides access to commercial software and data resources free of charge to research users, with financial support provided by the Provost Office and the Norris Medical Library. They also provide in-house workshops, training, and vendor tutorials. There are approximately 10 major packages provided to USC research users, with more than 1,200 registered users. The number of "seats" for each software package is determined based on demand and available budget, and varies from package to package. Given the large number of users, it is estimated that the cost saving in having the multi-seat site license as opposed to individual purchase is at least one order of magnitude. In general, it is estimated that for most software packages, acquiring a site license (over purchasing individual licenses) is financially worthwhile for USC if there are 10 or more individual users.

Finding 5 (Bioinformatics Program model):
The Bioinformatics Program model for providing software support to research users appears to be a successful one: they provide a wide range of software tools to a large number of users; no major gaps in coverage have been identified by them or by their users.

At present, this is the only university level resource integrating software, training and support from a single university unit. However, looking ahead, the Program will need enhanced computing resources (for example by enhancing their computing condo at the High Performance Computing Center, HPCC), additional seats for certain packages, tools for upgrading data security, and data storage capabilities in the next 3 years. These needs have not yet been quantified, but should be in the near term.

Finding 6 (USC Libraries):
USC Libraries constitutes the University's core expertise in archiving extremely large data sets. The USC Digital Repository, operated jointly with ITS, currently holds 94 petabytes of data with a goal of 2 exabytes of data. They support digital data for organizations such as Warner Brothers and the Shoah Foundation, among many others. To ensure no data are lost, USC Libraries refresh their media once every 3 years, as well as keep a mirror site at Clemson University. The USC Libraries expertise is in archiving image data, photos, sound, and music; they specialize in archiving collections and content, not providing or supporting software. They are currently working on getting HIPPA-B certification and approval to support health-related data and imaging archives.

Finding 7 (USC Libraries data archiving funding):
USC Libraries carries out its work primarily through contracts (e.g., Warner Brothers and Shoah Foundation). Any data archiving, hosting, metadata generation, web page maintenance, etc., needs to be supported by individual projects (i.e., faculty funding). Also, the repository is not currently conceived as a resource to enable broad access to USC generated data, but is instead functions as a location for long terms storage of data for individual projects.

Finding 8 (Dissemination of information regarding computing resources):
The URC noted that in many cases, it is difficult to acquire information about many of the above resources. This especially applies to determining which software is available and/or supported by USC, and has resulted in faculty making individual software purchases through independent funding sources. One example is Endnote, a frequently used program that is individually purchased by many faculty members despite the free availability of this software through the USC Norris Library. This is partially because of incomplete information in websites and other outlets, and partially because the computing, data handling, and software resources are scattered among several different entities, making an exhaustive search a difficult task.

Finding 9 (faculty survey regarding software):
The faculty software and database access survey had 445 respondents. A significant majority (70%) indicated that their primary software interest, whether already at-hand or desired, was under the "Statistics / Data Analysis / Data Capture / Data Visualization" category. Among the top 10 most popular software and database packages, only 4 (SPSS, STATA, MATLAB and SAS) are currently being supported by ITS. None of the next 10 most favored are being supported. Results are summarized in the appendix.

Finding 10 (Data storage needs within the USC research community)

Data storage presents a wide array of challenges in the research community due to the many types of data that needs to collected and stored within an academic institution, which can include both static information and software or processes needed for access. Although some of these data merely needs to be stored for record keeping, other data require sharing within USC, with other collaborating institutions, or even publicly (especially with the recent focus on "reproducibility" in research and sponsor requirements). Maintaining security of the stored data is crucial as one considers data storage solutions, and a number of different levels of security can be required for data storage in an academic institution. These include a.) classified data, b.) highly-protected (restricted) data including research for the Department of Defense (DOE) and the Department of Energy (DOE), health-related data (HIPPA compliance), and student information (FERPA compliance) c.) more general information, but categorized as sensitive and d.) already-published data (public).

Finding 11 (Current available data storage at USC):

At present, there are very limited research data storage options for USC faculty, and those options currently in place are not available broadly to the USC research community. The Shoah Foundation provides data storage and sells or leases data storage to USC at uncompetitive rates. The Center for High-Performance Computing (HPC) provides data storage and data analysis tools for a small number of high-powered users. There are no university-wide systems currently in place for providing the broad USC research community data storage options and a large proportion of USC faculty store their research data on their local computers, on resources coordinated within their units, on commercial data storage platforms (such as Dropbox), or occasionally through national / international research data repositories (such as Open Science Foundation, ResearchGate, Academia, etc.). Microsoft OneDrive is available freely under contract to USC faculty, although this information may not been successfully disseminated to a large proportion of USC faculty.

Finding 12 (Feasibility of cloud storage for USC research data):

In considering the feasibility of cloud storage as the predominant data storage solution for USC research, a number of advantages can be proposed (and compared to local data storage, including: 1.) lower cost, 2.) scalability, 3.) elasticity (the ability to purchase the amount required), 4.) rapid deployment, and 5.) potentially better fault tolerance. Advantages of local storage: 1.) security and privacy concerns (although these remain concerns with local storage, cloud storage may have higher risks in this regard), 2.) less dependency and vendor lock-in concerns (including the potential for vendor default), and 3.) more control of storage mechanisms. Although there are concerns with cloud storage of highly-protected information, such as DOD, DOE and HIPPA-compliance, the majority of large cloud vendors (including Amazon and Google) are now able to provide HIPPA-compliant storage that meet the large majority of USC's data-storage security needs. Classified data (including DOD) will likely continue to require local storage for the foreseeable future.

Finding 13 (Internal Audit Report on Data Backup and Disaster Recovery)

The July 2015 Internal Audit Report on Data Backup and Disaster Recovery from University and USC Hospital Operations specifically addressed research data backup and data protection issues. The Report observed that research data is not centrally managed by the

university and that schools and departments currently have loose oversight over research data. Researchers and faculty members therefore choose for themselves methods for managing their research data. The Report emphasized that research data may potentially contain sensitive information (such as personal health information covered by HIPPA) that has not been de-identified and that adequate controls are not sufficiently in place by the schools and departments at USC to protect or restore sensitive research data in the event of an emergency or disaster. The Report recommended that research data be managed centrally so that there is more oversight and control for protection of this data, but conceded that this recommendation may not be practical to implement. Therefore, if management of research data is to remain decentralized, a recommendation was made for a periodic communication to be sent to schools and departments to require close adherence to security and protection protocols regarding research data. In regards to data protection and compliance, a specific recommendation was made that schools and departments request that faculty members, researchers and students do not store sensitive information on department file servers without de-identifying the data prior to storage. In addition, it was reiterated that USC policy on *Information Security* states that all university data be reviewed on a periodic basis and classified according to its use, security and importance to the university (with general classifications as either restricted, sensitive or public).

Finding 14 (Varied academic computing strategies at USC satellite locations)
The data storage and sharing requirements of USC researchers and faculty span several satellite locations that are already engaged in similar efforts to optimize their solutions to these recognized data storage needs (including Children's Hospital Los Angeles, Institute for Creative Technologies, and Information Sciences Institute). These satellite locations employ USC faculty whose research priorities and obligations contribute to and impact USC. In addition, faculty at Children's Hospital Los Angeles work with staff members (including research assistants, research coordinators, nurses and other clinical employees involved with research) who are considered CHLA staff and are not affiliated with USC. Access to USC's computing resources (Pubmed access, software packages and data storage) is not provided to CHLA staff, despite the fact that many are involved directly on research projects with USC faculty members.

URC Recommendations

Recommendation 1 (unification of academic computing resources):
The URC recommends an effort to harmonize and/or unify academic computing resources at USC, at a minimum focusing on providing a centralized resource for information about available resources (see Recommendation 3). Software, database, data archiving, and computational resources are scattered among several different entities on campus. Furthermore, a large degree of variability/ is present among the various schools and departments in terms of availability and access to these academic computing resources.

ITS appears to have the broadest capabilities in this regard. In its Mission statement, ITS states that "The mission of ITS is to ensure the effective application and integration of advanced information technology in support of the teaching, learning, research, and administrative missions of the university. Clear leadership and guidance within ITS enable the organization to fulfill its mission in order to serve the entire USC community with distinction." This mission is in line with the computing needs of the USC research community, making ITS the most likely entity to lead such an effort.

Recommendation 2 (software purchasing decisions to align with faculty needs):
Results of the faculty survey suggest that faculty research software/database needs are not being met through ITS. A more direct input mechanism from faculty research groups is required in the ITS software acquisition decision-making process. To stay true to its mission statement and its key goals (http://itservices.usc.edu/about/), ITS needs to institute effective measures for seeking and implementing inputs from the entire USC community including faculty, staff, and students. While there are clearly software and database needs that are specific to individual departments/ schools, ITS should take the lead when these needs are shared across enough academic units. To help ITS achieve its goals and vision, the URC recommends that an advisory committee composed of these key stakeholder groups (including members of each of USC schools) advise ITS and evaluate its performance on an annual basis. The Academic Senate / Office of the Provost Committee on Information Services may be ideally suited for this role and should thus be considered for this task. A mechanism to more readily allow stakeholders to suggest desired software / database resources should also be implemented (e.g., an online submission form) in addition to an annual faculty survey to assess these computing needs.

Recommendation 3 (data storage for faculty):
The URC recommends that secure data storage and backup be provided to all USC faculty. There is tremendous variability in the types of data storage (and quantity) that are required by USC faculty, but a minimum amount of storage should be made available to all faculty. A cloud based solution appears most practical, with a minimum storage amount (e.g., 100 GB) made available to each faculty user. Some individual choice should be allowed regarding the type of data storage utilized, but a focus on fulfillment of standards should be emphasized. This includes meeting privacy and security standards (such as HIPPA) and appropriate sharing of data (including fulfilling expectations of data sharing as required by grants / publications). Although a centralization of research data storage has been proposed by the USC Internal Audit Report, an alternative strategy of a decentralized school-based strategy (following a single university standard) should be considered and may be most practical when considering a cloud-based storage approach. This broad university standard should include the following requirements: 1.) mandate encryption of data in accordance with University and funding agency policies, 2.) ensure that sensitive data is de-identified, 3.) require monitoring and periodic review that sensitive research data cannot be accessed by unauthorized users, 4.) provide personnel and technical tools to assist with these processes in order to motivate researchers to adhere to data protection requirements. 5.) ensure an infrastructure for the appropriate emergency back-up of this data. As part of these efforts, the URC recommends a coordinated university effort to

ensure compliance with external data regulations, including US Government (classified and other levels), medical (HIPPA), and student (FERPA). A significant commitment (both financial and through support personnel) will be required by the university to achieve these recommendations for data storage, but this effort appears crucial for the future of USC Research.

Recommendation 4 (data sharing capabilities for faculty):

In addition to the requirements for data storage (and reliable back-up), researchers and faculty are expected to comply with a number of data sharing requirements. Data sharing strategies are requirements for many grants / publications, and a recent focus on these requirements has been seen due to a broad academic emphasis on "reproducibility" of research. The URC recommends that tools and resources be provided to researchers and faculty in order to facilitate sharing and access to data, both for the unique needs of the researcher and to comply with federal requirements of data sharing. A number of scenarios are envisioned regarding research data sharing at USC and require individual solutions / approaches: 1.) large projects with USC as the lead data repository, including both sponsor-funded projects (with clear cost-recovery mechanism) and data repositories initiated by USC as strategic investments (e.g., Shoah Foundation data repository), 2.) projects in which USC participates in community storage, but does not lead the repository; and 3.) smaller projects that required individual data storage solutions outside of established community-sharing repositories. Each of these require unique tools and resources, but in many cases, current federal policies require that solutions be in place for research to proceed at the university. In cases of sponsor-funded research, a cost-recovery mechanism may be in place during the funded period, but strategies for data retention beyond the end of a funded agreement are often required by sponsor agreements. In addition, creation of innovative data repositories with strategic goals for the university provide an opportunity to further USC's academic reputation and should be encouraged and prioritized by the university.

Recommendation 5 (information dissemination regarding academic computing):

Although ITS and other computing units provide information regarding their available resources in a number of online locations, these many sites are difficult to navigate for most users. A centralized webpage should be created to orient USC researchers and provide links to available computing resources within the university. This would serve to help users parse through the multitudes of available resources and may also make the gaps in available resources more clear. This unified portal could 1.) enable more efficient searches for available site-licensed software, 2.) allow easier access to learning tutorials, 3.) provide a more readily-available mechanism for requesting staff assistance, 4.) allow a means for disseminating important / timely computing information to the USC research community, 5.) provide a forum for suggesting / requesting new software purchases, and 6.) provide guidance regarding options at USC for the various types of data storage and data sharing.

Recommendation 6 (Parallel processes for academic computing among satellite USC institutions):

The URC recommends that ITS and research leadership at USC and its satellite institutions (including Children's Hospital Los Angeles, Institute for Creative Technologies, and Information Sciences Institute) engage in parallel discussions regarding academic computing efforts in order to maximize efficiency and potentially minimize costs of implementation. More specifically, the URC recommends a more collaborative effort regarding academic computing resources between CHLA and USC (with a focus on allowing access to these resources for CHLA staff that support USC faculty). A consolidated academic computing relationship may not only be financially beneficial to both CHLA and USC, but may also result in improved research collaboration between the two institutions.