Management Approach

Our Overarching Vision for Data Sharing

The PI is strongly committed to reproducibility of all research claims. This requires the availability, in perpetuity, of all data/code created by the research (including data that does not make it into the published findings, to reduce the file-drawer effect[1]). PI has exemplary track record in sharing data. Consider the following:

- Since 2008, PI Mueen has released all data and code while a paper is under review (rather than waiting for the paper's acceptance) thus allowing reviewers to reproduce experiments when reviewing the work.
- PI Mueen has started an extensive effort in collecting and sharing real datasets from internet (e.g. Dictionary data from Oxford Dictionaries), industry (Review data from TripAdviser.com and Hotels.com) and research labs (e.g. insect telemetry and human activity data).
- To ensure reproducibility, PI is committed to make the experimental results, parameter settings, executable, video demonstration and presentation slides publicly available.

This project will continue this policy of openness, sharing, reproducibility and transparency. For anyone that wants access to the entire dataset, we will give access to our database within capacity and, in general, ship flash drives. We will make the data available online in smaller subsets to facilitate separate research in spatial, temporal and graph mining.

Policy and Practice

Good data management requires computing infrastructure and access to expertise in archival management. Both are integrated into the UNM's policies and practices on data management. As the lead institution for this project, we will encourage our collaborators to provide their data for dissemination together with ours. They also will be free to make the data available through their own institutional or individual resources.

All principal investigators are provided with sufficient computer storage to store and process the data generated in their research. Every principal investigator has the ability to create his or her personal web site, and to make data sets available through that web site. Complementing this infrastructure, the Library Facilities are available to principal investigators to consult on the formatting of data from project inception through dissemination of finished products.

[1] Ioannidis J (2005). "*Why most published research findings are false*". PLoS Med 2 (8): e124. doi:10.1371/journal.pmed.0020124. PMID 16060722

Scope

This Data Management Plan acknowledges that primary data "commonly accepted in the scientific community as necessary to validate research findings" be made available at little or no cost to the PI or project. In accordance with this policy and guidance from the Office of Management and Budget, this plan does not include preliminary analyses (including raw data), drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. Data that must be withheld long enough to enable peer review and publication/dissemination or protection of intellectual property is subject to this plan only after those steps have taken place.

If applicable, the management of personal data about users or any other human subjects is subject to policies and restrictions in protocols adopted by the relevant Institutional Research Board (IRB) and the Family Educational Rights and Privacy Act (FERPA) regulations. When appropriate, we will use anonymization and controlled dissemination and exercise the utmost care in sharing even abstracted or aggregate level statistics.

As stated in the project description, we intend to generate course materials (ppts etc), scientific articles and Technical Reports, algorithms, and software that we intend to share with the community.

Data and Metadata Format and Content

All completed articles will be formatted to conform to the EPrints open access platform. This will build an important bridge between data sets managed at the laboratory or academic unit level and archives managed by experts in library and information science.

Accessibility and Data Protection

All applicable electronic materials will be made available through the EPrints server. As discussed in the proposal and in accordance to University police, data will be made available only after appropriate steps have been taken to protect intellectual property. Confidential material will be handled according to policies and protocols for human subjects, FERPA, and any other applicable regulations and restrictions.

The dissemination server will be read-only; therefore, data will be secure from tampering. Files will be stored on a secure file server and will be backed up for robustness.

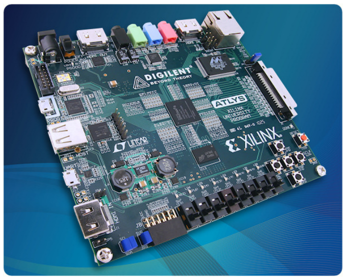
Derivative Products

All materials will contain acknowledgement of ONR support as per ONR policy.

In keeping with standard ethical practices, it is expected that subsequent users of the data will acknowledge the source.

Compute Resource

Dr. Mueen has been awarded a Microsoft Azure Research Award with 32 nodes in Azure with exclusive access. This resource will be the test bed for software techniques we develop in this project.



Dr. Mueen has two Atlys™ Spartan-6 FPGA Development Boards that are programmable through standard off-the-shelf desktop computers. UNM has licensed software packages necessary to synthesize circuits and transfer them to the board. These boards are portable and can be easily carried.

Hardware of special note includes: The Center for Advanced Research Computing (CARC) that supports researchers with cutting edge hardware support such as supercomputer, compute clusters, visualization tools etc. One of the systems named Poblano is an IBM P5-570 SMP system with 16 processors (1.9 Ghz) and 256 gigabytes of memory, which we will use for hardware accelerated algorithms. If needed, we can use the distributed systems in CARC such as the Roadrunner and Blackbear clusters.

Facility

The UNM Computer Science Department has over 200 general-purpose workstations and servers running Linux, Windows and OS X. The department network infrastructure consists of a switched 1GB backbone, which links the campus network and supports the principal departmental servers. Workstations connect via Ethernet ports, based upon hardware class and usage, including 336 10/100 ports and 154 Gigabit ports provided by a redundant stack of Cisco 3750 switches. The CS Department has 4 class C subnets, which are segmented, into smaller broadcast domains.

Departmental services such as LDAP, DNS, DHCP, Subversion and Apache are hosted on redundant Xen servers. There are also five 8-core compute servers available for additional processing support. There is a departmental lab available to students with 18 workstations, primarily Linux, a classroom with 55 Linux based workstations, as well as several labs dedicated to specific research projects. The Computer Science Department also maintains a wireless network (80211.b) in the Farris Engineering Center for faculty and students. Additional labs are being planned. Systems support is provided by two full time staff and one part-time student.