**DATA MANAGEMENT PLAN**

***Our Overarching Vision for Data Sharing***

We are strongly committed to reproducibility of all our 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]](#footnote-1)). We have exemplary track record in sharing data. Consider the following:

* Since 2008, PI Mueen has released all data and code for every paper he published. He also release data to anonymous reviewers at the time of paper submission (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, pre-processing and sharing real datasets from industry (e.g. the data shown in the proposal) and research labs (e.g. insect telemetry and human activity data). PI Mueen has recently contributed two datasets to the UCR time series archive published in July 2015.
* 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. As mentioned in the description, we have collected more than a terabyte of data and we expect to produce 500 GB of data each year. Our data is currently stored in a shared mysql server in the department of computer science which is also used for instructional purpose. We budgeted money for a powerful server to store all our data safely in isolation and without sharing storage and compute capacity.

We will differentiate between data and results for sharing because of varying policies of social media sites. Almost all the sites allow usage of their data for research purposes, while some do not allow redistribution. We will share *all* of our code and generated results and provide community support for reproducibility throughout the duration of the project. For anyone that wants access to a big chunk of distributable data, we will provide limited number of accounts to our systems as researcher and, in general, we will return flash/hard drive with data. We will make the data available online in *smaller anonymized subsets* for reproducibility of results. We will develop a web-portal for reporting abusive user accounts every day. The portal will also host an archive of past results for free download.

***1. 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 project personal have the ability to create their personal web site, and to make results available through their web site. Complementing this infrastructure, the Library Facilities are available to the principal investigator to consult on the formatting of data from project inception through dissemination of finished products.

***2. Scope***

This Data Management Plan addresses the NSF’s policy 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.

We collect only publicly available data and our data is not identifiable because social media sites are anonymous in the first place. We will use Amazon Mechanical Turk to perform evaluation of our results. An IRB application is in place for crowd sourced experiments. We acknowledge that 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, videos, etc.), scientific articles and technical reports, algorithms, and software that we intend to share with the community. All of these materials will be shared online for free,

***3. 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.

***4. 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 policy, 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.

***5. Derivative Products***

All materials will contain acknowledgement of NSF support as per NSF policy. In keeping with standard ethical practices, it is expected that subsequent users of the data will acknowledge the source.

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