Social and Decision Analytics Division Data Management Plan

November 15, 2018

The Social and Decision Analytics Division in the Biocomplexity Insitute & Initiative of the University of Virginia unites statisticians and social scientists to create quantitative methods at a scale able to leverage ***all*** data to support policy and decision- making under uncertainty. SDAD is advancing statistical and quantitative social science methods to integrate novel sources of data with traditional sources of data to obtain a more comprehensive understanding of social problems. To do this, SDAD handles a wide range of data for their research projects. Many of the data sources are publicly available or **open-source**. Other data require data sharing agreements with the owner of the data. Examples of **data owners** are our sponsors, government agencies, or private sector vendors who acquire, repackage, and sell data. A **data-sharing agreement** is a formal contract that clearly documents what data are being shared, how the data can be used, and the requirements for maintaining the security and confidentiality of the data.

Before conducting research, SDAD submits and obtains approval from the appropriate Institutional Review Board (IRB). Data are managed to meet the requirements set forth by data data owners and outlined in the IRB protocols. These requirements demand coherent and effective operating procedures for the management of data that include transfer, storage, access, dissemination, and destruction processes.

- **Data Types**

  The data types are expected to include electronically-collected administrative and survey data, physically gathered aggregate field data, and individual-level data records (e.g., statute-protected, statute non-protected, and vender requirements). The SDAD team employs a data management plan that ensures personal identity protection and tracking of data provenance.

- **Data Transfer**

  All electronic data transfers will be performed via encrypted means. Supported network protocols for remote data transfer will include FTPS, SFTP, SCP, and for public-facing web applications, HTTPS. Data sets that are to be transferred manually (not via remote network connection) will be transferred using encrypted USB storage devices employing, at a minimum, an EncFS-based encrypted file partition.

  Email is not considered secure and will not be used to transmit protected data unless additional file-level encryption tools, requested and approved by the data provider, are employed.

- **Data Storage**

  Data stored “at rest” on centralized servers will be protected with filesystem-level encryption. Logical Volume Management (LVM) partitions on partner institution servers will be encrypted using Linux Unified Key Setup, LUKS. This setup is typically referred to as “LUKS-on-LVM”. LUKS is a disk-encryption specification that is based on an enhanced version of cryptsetup, using dm-crypt as the disk encryption backend. Direct access to data sets will be restricted to SDAD project-approved personnel and data managers. The method of access for loading and management purposes is via SSH using RSA encrypted key pairs.

  SDAD data management personnel will perform regular backups of the encrypted volumes via LVM snapshots.

- **Data Access**

  Data provenance and management best practices are included as part of project team training. By policy, copying of protected server-hosted data to project staff and student workstations will be disallowed—work will be completed via encrypted remote access to servers, both for software development and data analysis.

  Project staff are trained on best practices in data management, including 1) archiving original research data in a read-only format, 2) restricting data access to authorized users or personnel, 3) creative commons licensing and 'open access' data generation and use where appropriate, 4) proper password and authentication management in the case of public-facing services or applications (e.g. storing only hashed and salted passwords), and 5) legal issues concerning data gathering and consent (such as EU GDPR data protection regulations).

  Access to and management of data records is defined in consultation with stakeholders and the IRB. In general, originally provided data records are accessible via secured remote access only in a read-only format to authorized users. The intention is that original data sources are never modified. Resulting modified read-write data that are produced from the original data sources are stored back to a secure server and only accessible via secured remote access.

  For projects involving protected information, unless special authorization is given, researchers do not have direct access to data files. For those projects, data access is mediated by the use of different data analysis tools hosted on their own secure servers that connect to the data server via authenticated protocols. Results of the analyses of protected data are only able to be saved to the researcher's home directory on the hosting server. Researchers are not be permitted to save analysis results to their office workstations.

- **Data Dissemination**

  Data that are eligible for public dissemination will be disseminated through technical reports, peer reviewed publication, and open source webpages. Those data sets will include relevant metadata and provenance.

  Data that are eligible for private dissemination follow the protocols set forth in the Data Transfer section above.

- **Data Destruction**

  As required by specific data use agreements, completed project data are either securely archived on encrypted backup servers or destroyed. For data that are to be destroyed, the encrypted LVM partition on which the data are housed is wiped using the Unix 'shred' utility, overwriting all partition data files three times with multiple patterns by default. After the data are overwritten, the encrypted LVM partition will be destroyed.