Data Management Plan:
The New Mexico SMART Grid Center: Sustainable, Modular, Adaptive, Resilient, and Transactive

Types of Data Produced

SMART Grid Center participants will create research, training, and outreach products. **Research products** include scientific papers, technical reports, source code, hardware schema, data classifiers (e.g., to detect disturbances and faults), and environmental and electrical usage data. **Training products** will consist of online course modules and complete curricula in areas relevant to smart grids. **Outreach products** will include reports, lay or popular articles, and print or online materials promoting awareness and knowledge about smart grids (e.g., presentations, exhibit materials, and educational toolkits). Documents and training and outreach materials will be generated using LaTeX, MS Word, MS PowerPoint, and Adobe products. Research data products will include results from simulations, observations, and data collected from testbed deployments, program code (e.g., GridLAB-D, C++, R, MATLAB, Python), climate data, household electricity data, and survey data. Test and measured data from different sources (e.g., sensors, lab instruments, PMUs) will be stored in native file format, and then converted to preservation-ready formats (e.g., csv, txt) for archiving purposes. Data that underlie the findings reported in a journal article or conference paper will be deposited in accordance with the policies of the publication, or in public repositories (e.g., IEEE DataPort). Students will receive training in the use of laboratory notebooks (e.g., Jupyter Notebook, http://jupyter.org/) and best practices for QA/QC and data preservation and sharing. Data will be stored on laboratory and institutional servers and backed up at regular intervals to offsite storage. Co-investigators will periodically review the integrity and the quality of the data to ensure that the digital data is of high quality and is properly organized and maintained.

Data and Metadata Standards

Test and measured data from different sources will be stored in native file formats and converted to non-proprietary text formats for archiving purposes. Simulations and codes will be a combination of specific software types (e.g., .mat for MATLAB) and .txt files, and they will be stored in a version control repository such as GitHub. Some tabular data collected will be captured in spreadsheets or data tables and saved in .csv files for long-term accessibility. In addition, some data will be stored within a relational database (e.g., MySQL or PostgreSQL), within which metadata and documentation including database schema, query, and table definitions will be captured along with tabular data using SQL.

Minimally, information necessary to record the provenance or chain of custody and attribution of data will be documented using the Dublin Core metadata standard (DCMI). Information including dates, the names of contributing investigators and their organizations, as well as pertinent geospatial and descriptive information, will be labeled using the DCMI and recorded in README files stored or published as separate metadata document(s) with the data. Annotations using more advanced metadata (e.g., formal ontologies) will be provided where possible (e.g., using the ENERsip ontology).

Strategies for capturing metadata, and the structure and file formats of metadata will adhere to accepted standards for the different data types described above. Standards will be selected because they are community based, receive widespread usage, and are relevant to the discovery, unambiguous interpretation, and integration of the particular data type.

Policies for Access and Sharing

Data and software products generated by the project will be assessed for sharing and preservation using criteria of reproducibility, uniqueness, association with publications, and in the case of simulations, the cost of regeneration vs. preservation of outputs. Data and software products determined to be appropriate for sharing and preservation based on these criteria will be made available for discovery and reuse by researchers, students, policy makers, and the general public through publicly accessible repositories (see *archiving and preservation below* for a discussion of specific repositories) that enable robust discovery and access for the project's products. In order to enable publications based on the data and software products that are created, data and software will be subject to an embargo period of *no more than 12 months* from when they are produced, after which these products and their associated metadata will be placed in an appropriate repository. Upon submission of papers based upon specific project data or software, those products will be made available to reviewers while the paper is under review, and when published will be made available for public access through publicly accessible repositories for discovery and reuse, even if those products have not yet met the maximum 12-month embargo period. The only products for which this sharing policy will not apply are those that are limited in their redistribution due to license restrictions, privacy/human subjects, or are related to patentable technologies. In these limited cases a written waiver from the project data access and sharing policy must be obtained from the project PI that provides a justification and a plan for release to the extent allowable. We will follow University of New Mexico IRB protocols with respect to human subject data collection and accessibility, including anonymization of human subjects.

Policies for Re-use, Redistribution

Permissive licenses that maximize the potential reuse, modification, and sharing of data and software developed as part of the project by students, researchers, policy makers, and the general public will be standard for the project. In the case of data, the Creative Commons Attribution 4.0 (or later - https://creativecommons.org/licenses/by/4.0/) license will be applied to those products. This allows for sharing and adaption of project data products while requiring attribution to the creator(s) of those data. Software products generated by the project will be shared using the Apache 2.0 (or later - https://www.apache.org/licenses/LICENSE-2.0) license. The Apache 2.0 license is a permissive Open Source Initiative (OSI - https://opensource.org/) approved license that allows for a wide range of uses while also requiring maintenance of license and copyright notice and documentation of changes made. Any exceptions to these terms under which project produced data and software are shared must be obtained in writing from the project PI.

Plans for Archiving and Preservation

NM SMART Grid investigators are committed to the accessibility and preservation of the project's data products beyond the completion of the project. To the extent that data archiving may be subject to storage, networking, and other constraints, identification and selection of data products for archiving and preservation will be based upon criteria which emphasize research reproducibility, including the completeness and accuracy of data and associated metadata, uniqueness of the data and collection methodologies, and whether data are associated with published findings. Human subjects research will be de-identified, secured, and archived according to procedures and policies described in the corresponding IRB protocol.

Data and information selected for preservation will be maintained, curated, and archived in trusted repositories such as Dryad (http://datadryad.org/), figshare (https://figshare.com/) or other appropriate disciplinary repositories as listed within the Registry of Research Data Repositories (http://www.re3data.org/). Software code published in GitHub (https://github.com) will be assigned DOIs and archived in Zenodo (https://zenodo.org/). In cases where domain or discipline specific repositories are not available, data will be archived for a minimum of 5 years at the University of New Mexico (UNM) Libraries' Digital Repository (http://digitalrepository.unm.edu/) after the grant ends. The UNM Digital Repository is an Open Archives Initiative (OAI) compliant repository, which enables Dublin Core metadata and dataset objects to be shared and harvested by other archival and discovery systems through the OAI-PMH protocol.

Data archived within the UNM Libraries' Digital Repository will additionally be copied to and monitored within the Libraries' long-term preservation platform, LibSafe (http://www.digitalpreservationsoftware.com/digital-preservation-solutions/libsafe-digital-preservation-software/). The LibSafe platform provides National Digital Stewardship Alliance (NDSA) level 4 preservation, including geographically distributed replication, fixity checking, and format validation and migration. Ahead of ingest into archival repositories or LibSafe, development of preservation metadata using the Metadata Encoding and Transmission Standard (METS) and Preservation Metadata Implementation Strategies (PREMIS) standards will be coordinated with the Libraries' Data Curation Librarian. The UNM Libraries' LibSafe instance extended capacity for the proposed research includes up to 10TB of project data, with secondary preservation provided through the Libraries' Digital Preservation Network (DPN) membership. The combination of LibSafe and DPN affords 20+ years of replicated, dark archival preservation for archived project data.

After 5 years, disposition of data preserved within LibSafe will be appraised per established collection and archival management policies.

Data Privacy and Confidentiality

The project team has developed the following policies and practices to address management of human subjects research, IoT data streams, and other data with privacy or confidentiality risks. These policies are applied at relevant stages across the data lifecycle:

Data Collection

Specific to the collection and safeguarding of human subjects data via surveys and/or experiments, we follow the standards set forth by the University of New Mexico's Institutional Review Board (IRB)[1]. Of paramount importance is that we collect only the minimum necessary subject identifiers. In addition we safeguard individual data by severing it from the rest of the data set during our analysis period; we destroy any potential subject identifiers as soon as they are no longer needed; we limit physical access to an area or computer device where subject identifiers are stored; we encrypt data stored on portable devices; data with subject identifiers is not stored on public online or cloud storage services; and data stored off campus are password protected. Survey data collected by Qualtrics is de-identified when received, but, for consistency, is stored similarly to that above. Qualtrics adheres to industry standards for data collection and storage[2]. Participants are provided with information concerning data confidentiality and security in the informed consent.

Data Storage and Analysis

NM EPSCoR SMART Grid researchers are currently developing a streaming dynamic database which will store and stream real-time observations made at sensors on the grid. We adopt standard security features for various usages of this database.

1. Communication security: We adopt standard HTTPS protocol to stream the data over the internet. Hyper Text Transfer Protocol Secure (HTTPS) over Secure Socket Layer (SSL) is industry standard.
2. Data access control: Role based access control will be implemented to grant access to project personals at various privileges. Senior personals and PIs will be granted full access to the data. Student researchers and guests will be granted access to needed segments of the data.
3. Storage media security: As per the ISO/IEC 27040 standards, we will use 128 bit encryption on stored data. All encryption processing (such as key generation and activation) will be properly logged to enable security auditing on the encryption activities. Encryption and decryption will be hidden to the end users. In case encryption is an hindrance to performance (of frequent access), we will resort to an unencrypted database with strict access control to a very limited number of personals.
4. Code level security: We will maintain access and permission control on scripts and program files strictly to guard against attacks such as SQL injection attack and sniffing attack.
5. Disk sanitization and disposal: System administrators will properly dispose disks, and will zero-fill disks before recycling.

Data Sharing and Preservation

Selection and preparation of human subjects research and IoT device data for sharing and preservation will be guided by two principles: To protect the confidentiality of research subjects; and to maintain the integrity and research value of the data to support meaningful analyses and research reproducibility.

[1] The standards documentation is available at https://irb.unm.edu/sites/default/files/UNM%20Human%20Research%20Data%20Security%20Standards.pdf.

[2] Qualtrics is General Data Protection Regulation and California Consumer Privacy Act compliant; uses Transport Layer Security (TLS) encryption (also known as HTTPS) for all transmitted data, and; is both FedRamp and ISO 27001 certified. (For more information, see https://www.qualtrics.com/security-statement/).

It is important to note that many selection and preparation decisions may be determined by policies and procedures described in project specific IRB protocols. It is recommended that plans for data sharing and preservation be discussed with research subjects and included as part of the informed consent process. Data provided by third parties such as utility companies may also be subject to limitations specified in relevant data use or licensing agreements.

Data that can be published and preserved will be evaluated and processed in order to identify and manage direct and indirect identifiers. Direct identifiers are variables that uniquely identify subjects, and may include for example social security numbers, addresses, and driver's license numbers. Prior to publication and preservation within public access repositories, all direct identifiers will be removed from the data. Indirect identifiers are variables which are not sufficient in themselves to uniquely identify a subject, but which may be used to identify individuals through combination with other variables or information. Examples of indirect identifiers include zip codes, names of degree granting institutions and graduation years, and income. Indirect identifiers may possess significant value for data aggregation and analysis, so disclosure risks for these identifiers will be evaluated and addressed on a case by case basis. Where it is necessary and possible to do so, indirect identifiers in data for public access and preservation will be treated using methods including but not limited to removal, combining variables, top-coding, etc.

There will be cases in which it is not possible to fully de-identify data and retain its research value or enable reproducible analyses. These data will not be made publicly accessible or preserved in public access repositories, or in repositories that do not enforce appropriate access controls. Instead, such data will be evaluated for transfer to repositories that enforce variable levels of access control. For example, the Inter-university Consortium for Political and Social Research (ICPSR) offers different access levels to member institutions and researchers and supports access to sensitive data through the development of data use agreements and enforcement of physical and virtual data enclaves. The University of New Mexico is a member institution, so these ICPSR services are available to NM EPSCoR researchers.