Example Data Management Plan

An example of a DMP that would be appropriate to include in a 5-year project plan according to the guidance in P&P 630.0 follows. It was adapted from a DMP written on April 24, 2018, by Laurence Parnell of the Jean Mayer Human Nutrition Research Center on Aging (HNRCA), for a review of National Program 107 by OSQR.

# Data Management Plan

Data will be collected, processed, organized, housed, and retained based on standards appropriate to the biomedical research community guidelines. The greatest amount of data used throughout all aspects of this proposal are human population/cohort data. Personally-identifiable information (PII) is protected as part of Institutional Review Board approval; all volunteer and participant data we use is de-identified.

# Expected Data Types

Data that we collect, generate, or derive will include phenotypes/human traits/human laboratory variables, diet, SNP genotypes, CpG methylation epigenotypes, and microbiome data. Other data to be used include gene expression (from GEO NCBI and dbGaP), biological pathways (from KEGG, Reactome, WikiPathwys, various specialty), metabolites (FooDB, HMDB, Metabolon, Metabolomics Workbench) and their enzymes (enzyme database at Expasy), protein-protein interactions (BIND, BioGrid, STRING), human genome (Build 38, Ensembl), SNP-phenotype associations (GWAS catalog), CpGs (DNA methylation from Illumina annotation), eQTL and meQTL (GTEx/Jaffe 2015), cardiometabolic GxEs (CardioGxE), and various data from AHA Precision Medicine Platform and BioStudies repositories.

# Data Formats and Standards

Most data will be in flat file/text/csv or other open-file formats. Some data will be in Microsoft Excel (.xls), Portable Document Format (.pdf), and Joint Photographic Experts Group (.jpg) format. Laboratory protocols and experimental conditions will be produced in Microsoft Word, laboratory notebook, manufacturer’s pamphlet, or pdf format. When published to their respective data repositories phenotype/traits/laboratory variables, SNP genotypes, and diet data will be in flat file/text/csv format, as appropriate to the repository. Metadata and data standards appropriate to the biomedical research community guidelines such as FAIR, NIH’s Common Data Elements (CDE), and Minimum Information for Biological and Biomedical Investigations (MIBBI) will be used.

# Data Storage and Preservation

A research computing cluster at Tufts University that exceeds the project’s computational and storage needs is readily accessed over secure connections by laboratory members. Specific procedures regarding logs, firewalls, networked hardware, and/or security testing are all overseen by the Scientific Computing group of the HNRCA. Based on the amount of data we anticipate generating and the types of analyses to be performed, the estimated need is 50 TB of data storage (currently using 22 TB) and 1.0 TB of RAM (currently using 0.5). We will review these numbers annually in order to stay abreast of our group’s needs. LabArchives digital laboratory notebook software, via a license with our university partner, will be used to help store and organize laboratory protocols and research data. LabArchives provides for standardization of research practices and protocols, provides a complete and lasting record of discovery, and adheres to NIH and NSF research data management requirements.

Internal policy of the HNRCA meets or exceeds data retention and backup policy and guidelines of both NIH and NSF; we will adhere to this policy, which requires that all research records and data will be maintained for a minimum of 10 years after the termination of the research project (some minor exceptions where certain sponsors may require longer archival periods). All data will be stored on departmental network drives. These network drives have a concurrent backup system and are backed up daily to an off-site disaster recovery server. Additionally, weekly tape backups are performed. These multiple forms of storage and backup provide the redundancy to protect the data against any damaging event, including loss or degradation.

# Data Sharing and Public Access

Data acquisition, data handling, and electronic communication are carried out using desktop computers networked and linked to the central HNRCA server. Data acquisition (i.e., acquiring processed, de-identified data) will be via secure ftp. Data acquired from collaborators will not be disseminated to third parties, in accordance with standard agreements between collaborators. Data acquired or copied from public databases will retain their original source license/access designation. There will be no permission restrictions placed on data generated from this project. No issues should arise from our intention to share data publicly and data does not intend to be withheld.

Data will be made available upon publication via the following mechanisms. Per NIH guidelines and requirements of receipt of NIH funding, the PI must have a plan in place for management and deposit of population/cohort data. These data will be deposited into dbGaP (NIH) under the direction of the principal investigator. A substantial portion of our research is via the CHARGE Consortium (Cohorts for Heart and Aging Research in Genomic Epidemiology), and summary results from genomic studies are to be loaded into the CHARGE site of dbGaP after publication of the main findings. Outside of CHARGE, any data relevant to new clinical trials, including meta-analyses of existing population/cohort studies, will be deposited at clinicaltrials.gov. Although much of the data that will be used is already publicly available data that we intend to download and use for our analyses, we plan to deposit new data generated from this publicly available data in the repository of origin, with most data going to dbGaP. Microbiome sequence data will be deposited in GenBank. The Metabolomics Workbench (NIH) is a public repository for metabolomics data. Such data may be embargoed until either the completion of the study, which is standard practice, or until publication of the scientific report.

For data that are not appropriate for dbGaP or other disciplinary repositories, we intend to use the USDA National Agriculture Library’s Ag Data Commons as a repository. When applicable, we can submit a Data Descriptor article to the journal *Scientific Data* (Springer Nature) for peer review of data deposited in Ag Data Commons (data.nal.usda.gov/).

# Roles and Responsibilities

The principal investigator and the senior author(s) of a study have the responsibility to make certain that all non-PII data are made publicly available. The Biostatistics and Data Management Core Unit of the JM-HNRCA has responsibility for evaluating and assessing individual laboratory data management plans, and advising on and assisting with data acquisition, storage, and sharing responsibilities. When key personnel leave the project, the principal investigator will ensure the successful reassignment of project data management responsibilities.

# Monitoring and Reporting

The principal investigator and senior author(s) will revisit this data management plan, with the assistance of the Biostatistics and Data Management Core Unit of the JM-HNRCA if needed, on an annual or more frequent basis to make any necessary changes or additions to the plan. Data publications will be listed in annual project reports.