## Data Management Plan

The BAOBAB calibration, imaging, and data-reduction pipeline will be implemented using the open-source AIPY software framework, and will run on a 128-core, 120-TB computing cluster and RAID storage file system at the University of Pennsylvania. This cluster will also host the principle storage of BAOBAB data, and data distribution will be managed through account access to this system. All observed data will be stored and archived indefinitely at this location.

### Data Storage and Format

Data are recorded in the field to Redundant Arrays of Independent Disks (RAIDs) built from off-the-shelf disks. These disks will be removed and shipped to the computing center at the University of Pennsylvania as they fill to capacity. Upon arrival, data are copied again into a high-performance RAID system for access within the computing cluster. RAID has become a simple commercial commodity that provides both speed and reliability at comparatively low cost. Consumer grade disks are well-suited to field and shipping conditions where many storage units are desired, and where data redundancy permits replacement in the event of a failure. All data will be checksum-validated before the transport RAID arrays are erased and sent back to the field. The dominant cost of the field and transport system is the number units required by the latency of the transport system. This proposal budgets for 6 field storage disk arrays that will be used for recording the raw data output of the BAOBAB instrument.

The PAPER computing cluster at the University of Pennsylvania employs a parallel file system using the IBM General Parallel File System (GPFS). This system has the advantage of scaling linearly with volume through the target volume of 120 TB at a relatively low cost per gigabyte. Full correlation of 49 antennas with 1024 frequency channels integrated every 10 seconds generates approximately 170 GB of data per day, corresponding to 31 TB for a 180-day integration. This data volume can be accommodated on the existing 120-TB filesystem at the University of Pennsylvania.

Raw and calibrated visibility data will be recorded in the standard MIRIAD UV data format. Various open-source software tools are available from within MIRIAD, CASA, AIPY, and other software packages for converting this format into other commonly-used formats. The MIRIAD UV file format naturally includes all appropriate metadata for interpreting measurements. Calibration parameters are included into this format as they are derived. Intermediate and final data products such as images and sky maps will be stored in the standard FITS file format and HEALPIX-FITS format (for spherical data sets) used in astronomy and cosmic microwave background research.

### Software

BAOBAB analysis will be structured around the open-source AIPY software toolkit. This software is distributed through online repositories with full revision control. Additional general-purpose software developed for BAOBAB will be distributed through these same repositories. Internally, data analysis routines will be kept under revision control and distributed through a password-protected website. These routines will be made available to collaborators to the project.

### Access and Distribution

Data distribution will be managed through account access to the computing cluster and data repository at the University of Pennsylvania. All observed data will be stored indefinitely and made

available to collaborators. Data may also be distributed beyond the collaboration as requested. All data will be available by request to the community one year after observation, which is the amount of time necessary to ship data to the central repository, properly calibrate it, and include all appropriate metadata.

### Archival

The 120-TB storage system located at the University of Pennsylvania will be internally redundant, and furthermore will employ two completely redundant backup systems to store the full data volume (at slower access speeds). Data will be stored indefinitely on this system.