Data Management Plan
Extended Baryon Oscillation Spectroscopic Survey

Experiment description:

eBOSS is the cosmological component of the fourth generation of the Sloan Digital Sky Survey (SDSS-IV) located at Apache Point Observatory (APO) in New Mexico. Over the period July 1, 2014 through June 30, 2020, eBOSS will use wide-field spectroscopy from the 2.5-meter Sloan Foundation Telescope to obtain percent-level measurements of Baryon Acoustic Oscillations (BAO) from $0.6 < z < 3.5$. Using more than 1,000,000 spectra from galaxies and quasars as tracers of the underlying density field, eBOSS will probe the largest volume to date of any cosmological redshift survey. With four classes of spectroscopic target, eBOSS will enable the first high precision distance measurements at the epochs when dark energy emerged as the dominant dynamical component of the Universe.

DOE's roles in the experiment:

eBOSS will use the DOE-funded BOSS spectrographs that were proven in SDSS-III and made the best distance measurements to date at $z<0.57$ and $z\sim2.5$. The spectrograph design remains well-suited for the number density and luminosity of the galaxy and quasar targets that eBOSS will observe to explore entirely new epochs of cosmic history. The eBOSS survey is supported in part by DOE operations to provide a portion of the salaries for the technical staff who perform the observations each night.

Partnerships:

There are no partnerships with other federal agencies. As of this writing, DOE is the sole provider of federal funding for this project.

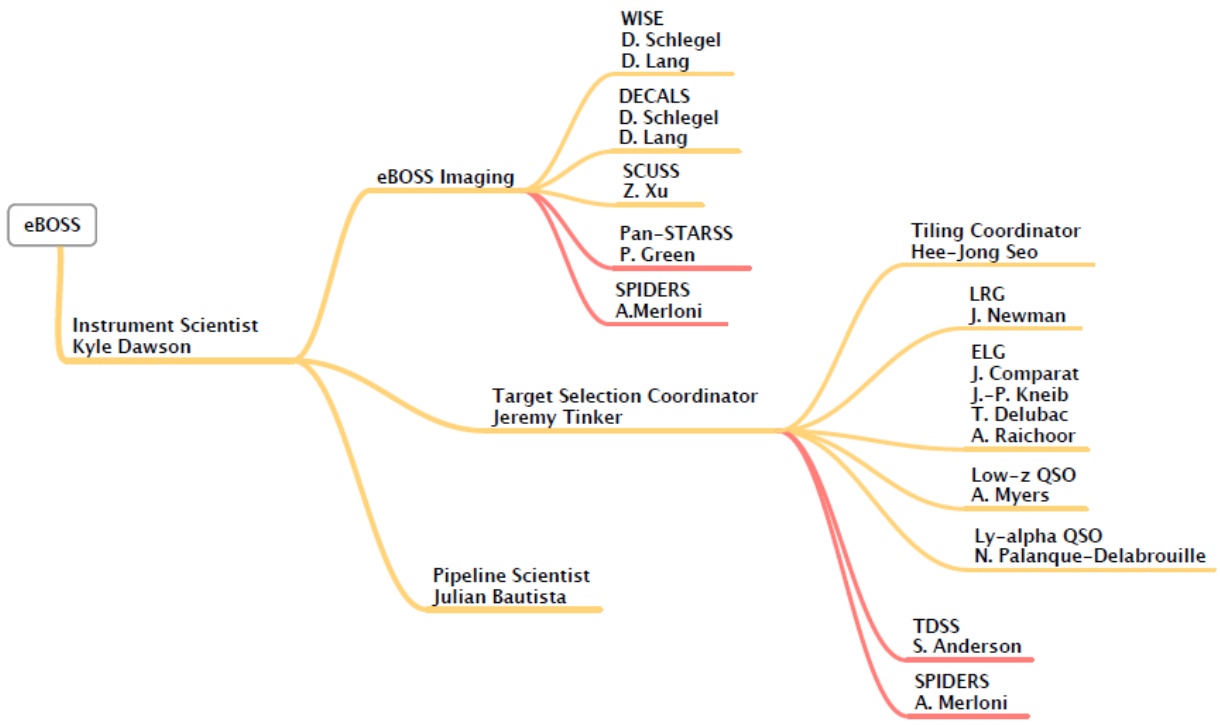
Organization – Agency/Lab level

eBOSS is one component of the SDSS-IV project. SDSS-IV is managed by the Astrophysical Research Consortium (ARC), whose members are Georgia State University, the Institute for Advanced Study, Johns Hopkins University, New Mexico State University, the University of Colorado Boulder, the University of Virginia, and the University of Washington. Funding for SDSS-IV is provided by the Sloan Foundation and a group of over 50 Participating Institutions. An Advisory Council consisting of representatives from these institutions serves as the governing body for the project. The operations management of the APO facility is the responsibility of New Mexico State University. The major computing for data processing and storage will be maintained at the University of Utah. Data reduction for eBOSS will be performed at the University of Utah. The science collaboration includes eleven institutions that have DOE Cosmic Frontiers funding.

Organization – Experiment level

The Principal Investigator (PI) of eBOSS is Jean-Paul Kneib at École Polytechnique Fédérale de Lausanne. The Instrument Scientist is Kyle Dawson at University of Utah (PI of the DOE operations grant). The Survey Scientist is Will Percival at University of Portsmouth.

The Instrument Scientist is responsible for the implementation of the eBOSS survey. He oversaw survey readiness in terms of targeting before the beginning of observations. He ensures that the survey progresses towards the completion of its goals and requirements now that observations have begun. Furthermore, he will monitor instrument and hardware performance throughout the survey. The organization of imaging, target selection, and pipeline effort is shown in the chart below:



The entire targeting effort is coordinated by Jeremy Tinker (NYU). As the Targeting Coordinator, he works with the targeting groups to ensure consistency of input file formats, verify that appropriate imaging data products are included in the internal and public releases of data, and determine what computing resources are adequate to store and access the imaging data.

Hee-Jong Seo (University of Ohio) is the Tiling Coordinator. She assigns fibers to LRG, ELG, and quasar targets once the selection from imaging data is complete. The primary responsibility of the Tiling Coordinator is to define the geometry of the survey by specifying the location of all fibers and plates to be observed. The tiling process provides the link between the targeting data (as inputs) and the final processed spectroscopic data (as the final outputs of the survey).

The eBOSS software pipelines will rely primarily on the well-developed pipelines for BOSS. Julian Bautista (Utah) is the eBOSS Lead Data Scientist and oversees this effort.

Collaboration:

The SDSS-IV collaboration has received financial contributions from 25 institutions who have signed as full members. There are six participation groups comprising 24 institutions of named individuals with data rights. There are 23 institutions who have signed as associate members, with data rights to a limited number of scientists. These members have full access to the eBOSS data and the data obtained from the other SDSS-IV surveys during the proprietary period. Those members of SDSS-IV who are specifically interested in eBOSS subscribe to an internal mailing list. There are 262 individuals subscribed to the eBOSS mailing list.

The collaboration policies are described in the SDSS-IV Principles of Operations and in the SDSS-IV Publication Policy. The elected Scientific Spokesperson leads the science collaboration with the support of the Collaboration Council, which consists of representatives from member institutions. There are a number of Scientific Working Groups which coordinate the efforts of collaboration members, organized according to broad scientific themes.

Data policy management:

The SDSS-IV Principles of Operations and Publication Policy define the policies for use and publication of data prior to its public release. The SDSS-IV Director (Michael Blanton, NYU) has final authority on determining when data can be released to the public. There is a dedicated Data Management team, led by Joel Brownstein (U. of Utah), Anne-Marie Weijmans (St. Andrews), and Ani Thakar (JHU) that has primary responsibility for data management, data documentation, and data distribution to the collaboration and the public.

Data Description & Processing:

The eBOSS survey produces 1000 spectra per 15 minute exposure, each covering nearly 7000 angstroms in wavelength at a sampling of roughly 1 angstrom per pixel. Each 15 minute exposure is processed immediately on-site using a fast, automated data reduction pipeline to check quality. A postdoctoral researcher at the University of Utah is responsible for this data reduction pipeline and for establishing eBOSS operational procedures at the mountain.

These raw data are sent from the site of the observatory to a Science Archive Server (SAS) system at the University of Utah each morning. When the transfer is complete, they are fully processed by an automated data reduction pipeline that takes roughly 10 hours to reduce a full night of data. The data reduction pipeline consists of two largely independent steps: extraction and classification. The extraction step from the raw CCD images results in wavelength-calibrated and flux-calibrated spectra. In the second step, the one-dimensional spectra are classified into object types and redshift and recorded in a catalog. The Lead Data Scientist for eBOSS is responsible for ensuring this data-reduction pipeline runs successfully and for providing information on the data content to the collaboration.

Data Products and Releases:

The raw data, extracted spectra, and catalog of spectral classifications are available to members of the collaboration within 24 hours of the observation under nominal circumstances. The data are stored on the SAS at the University of Utah, mirrored to a Science Archive Mirror

(SAM) facility hosted at NERSC, and backed up to tape through NERSC's high-performance storage system (HPSS). The data are available to members of the collaboration through a web interface, rsync, and Globus transfers. The Center for High Performance Computing (CHPC) at the University of Utah provides much of the infrastructure, network specialists, and the tier-3 data center that houses the SDSS cluster and file system. In addition, members of the collaboration can obtain affiliate login accounts on the servers at University of Utah to access the data directly on local filesystems.

On regular intervals of roughly two years, the data-reduction pipeline will be frozen and all raw data will be re-processed into extracted spectra and a catalog of classifications. All three levels of data will be placed into a dedicated repository. The results of cosmological analysis will be published on roughly this schedule using the sample of each dedicated release as the basis for the measurement.

Public release of the eBOSS data will be made through four scheduled data releases that contain the identical information found in the dedicated samples for cosmology measurements. There is a proprietary period for data in each release such that at least one year passes from the time of the observation to the time of the public release. The first release will occur in July 2016 and will include a representative sample of roughly 100,000 spectra taken at the conclusion of SDSS-III and in the first year of SDSS-IV. The first two years of spectra will be released in July 2017. The first four years of spectra will be released in July 2019. The full sample will be released at the end of 2020. Each of these public releases of eBOSS data will be executed as part of a coordinated project-wide SDSS-IV data release (DR), with the next release (July 2017) being designated DR14.

The primary public portals for SDSS-IV data are the SAS at the U. of Utah (described above) and the Catalog Archive Server (CAS) and SkyServer system at Johns Hopkins University (JHU). SkyServer is a feature-rich interactive graphical web interface to the SDSS data set. SkyServer provides simple form-based access to the catalogs, and informative "Explore" pages for each photometric and spectroscopic object, suitable both for education and for quicklook purposes by researchers. The CAS is a comprehensive SQL database of SDSS catalog data that can be queried either through a SkyServer web interface or through the scriptable CASJobs system. It allows uploads of user data and server-side manipulations, making it a powerful platform for efficient, sophisticated catalog analysis.

Plan for Serving Data to the Collaboration and Community:

All of the data accessible to the eBOSS collaboration and used in the primary cosmological analyses will be documented and included in the four public releases, thus enabling validation of results by anyone outside of SDSS-IV. In addition to access by the CAS and Skyserver described above, the data will be presented in the same format as that used by the collaboration.

As in SDSS-III, each data release will be preceded by a "documentation festival" that gathers a focused team from across the collaboration to complete the online documentation and tutorials necessary for the broader community to take full advantage of the scientific potential of the SDSS-IV archive. All pipeline code associated with release data will simultaneously be made available to the public through the SDSS-IV SVN server.

Plan for Archiving Data:

Along with all SDSS-IV data, eBOSS data will be archived on RAIDed spinning disk arrays on the SAS and SAM through the duration of the project. All raw data and publicly released reductions are also written to long-term tape archival storage through HPSS. Each successive SDSS phase has included the data from all previous SDSS phases as part of its active online archive, thanks to continuous improvements in mass storage technology that have roughly matched the integrated volume growth of the SDSS archive. Any future SDSS phases are likewise anticipated to maintain eBOSS survey data accessible online.

Plan for Making Data Used in Publications Available:

The SDSS-IV collaboration maintains a public webpage that is updated regularly. The webpage contains links and documentation for the data products associated with each data release. The webpage also documents publications produced by the eBOSS team, along with links to catalogs and other research data produced for the papers. In accordance with the DOE requirements for data management, these web pages make research data associated with key eBOSS publications open and machine-readable. Furthermore, the key software codes used in published eBOSS analyses will be release publicly through the SDSS-IV SVN server.

Responsiveness to SC Statement on Digital Data Management

This data management plan fully follows SC Statement on Digital Data Management.