SDSS-III Project Execution Plan

Version 1.1

February 5, 2009

This is version 1.1 of the Project Execution Plan for SDSS-III, submitted to the National Science Foundation on February 6, 2009. This document draws on broad input from members of the SDSS-III survey teams. The principal authors and reviewers of the document text, who are responsible for its content, are:

Daniel Eisenstein, University of Arizona, Director

David Weinberg, Ohio State University, Project Scientist

Bruce Gillespie, Johns Hopkins University, Program Manager

David Schlegel, Lawrence Berkeley National Laboratory, BOSS PI

Steven Majewski, University of Virginia, APOGEE PI

Connie Rockosi, Lick Observatory, SEGUE-2 PI

Jian Ge, University of Florida, MARVELS PI

Jim Gunn, Princeton University, Observing Systems Lead and Infrastructure Lead

Mark Klaene, Apache Point Observatory, Site Operations Manager

Donald Schneider, Pennyslvania State University, Survey Coordinator

Michael Blanton, New York University, Data Coordinator

Jordan Raddick, Johns Hopkins University, Education and Public Outreach Coordinator

Michael Evans, University of Washington, ARC Business Manager

Revision History

Version 1.0: Submitted to NSF on December 31, 2008

Version 1.1: Updated Chapter 2 with text on APOGEE-MARVELS co-observing, more precise definitions of metrics. Added Appendix A on forecasts against these metrics. Added WBS, schedule, and budget as additional appendices, contained in separate files from this Word document. Added data distribution milestones section. Assorted cosmetic changes.

13. Data Processing, Archiving, and Distribution

The previous development chapters have described the plans for pipeline development. This chapter describes the procedures by which the pipelines get run and data get archived and distributed, especially who is responsible for what.

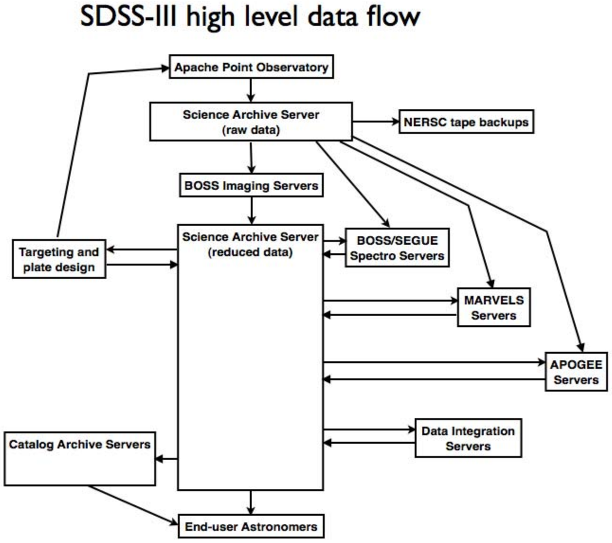

Figure 13.1: SDSS-III high level data flow diagram.

13.1. Data Archiving

The high-level data flow is shown in Figure 13.1 above. During observations, imaging, spectroscopic, and meta-data are written in real time stored temporarily in a staging area on an MJD basis. Checksums are produced at the time of observation. They are transferred to the SAS at LBL via high-speed internet connection. MJDs will only be deleted sometime after they have been archived and backed up to tape; a page on http://trac.sdss3.org conveys the status of each MJD for this purpose. Simultaneously, a second copy of the raw data is stored onto removable disks that are ultimately kept in offline storage at APO. The APO systems administration team (Jon Brinkmann and Fritz Stauffer) ensures that these steps are taken, in consultation with the instrument scientists.

Each morning after data for an MJD is taken, SAS retrieves the raw data and verifies its contents, checking for consistency among the files. It then backs the MJD up to the NERSC HPSS tape system. After each step (copy, verify, backup) SAS updates an internal database with the MJD status. A page on http://trac.sdss3.org posts this status, indicating after the backup step that the MJD is "ready to delete." In the evenings (APO time) the Science Archive Mirror (SAM) at NYU will update its mirror of SAS. The integrity of the data is checked through periodic checksums. The Data Archive Scientist is responsible for this process.

The individual teams download raw data from the SAS. In cases where the reduction pipelines depend on external datasets, the version of those external datasets used (2MASS, Tycho, GSC 2.3, SFD, etc) will be stored at the SAS and mirrored to the reduction institution, so that the pipeline codes have a reliable, consistent, and documented source for their inputs. In consultation with liaisons from each team, the Data Archive Scientist is responsible for maintaining these data sets and coordinating the downloads.

After data processing is complete for any portion of data (an imaging run or a particular instrument on a spectroscopic plate), the SAS will retrieve the reduced data. To facilitate this process, each reduction team will maintain a "rerun file" indicating the status of each re-processing of the data. Only those re-processings marked "complete" will be retrieved. Each set of reductions will be marked with version numbers indicating the software used. The survey teams will maintain a copy of their reductions, though not necessarily a copy of the raw data. When a re-processing is retrieved, it will be backed up in the HPSS system and also copied to the SAM. The upload and backup procedure will be maintained by the Data Archive Scientist.

Because of the timing of SEGUE-2 in the first year, it does not have access to the final reprocessing of the SDSS imaging data. Thus, for targeting and spectrophotometric calibration it relies on SDSS-II DR7 imaging catalogs. The relevant catalog entries used are stored in standard formats tracked in the data model with backups archived at SAS, SAM and HPSS.

Results from data assembly steps occurring at the SAS (window function determination, matching of spectra to photometry) will also have tracked versions, and be backed up at SAM and HPSS. The Data Archive Scientist is responsible for these backups.

Allowed file types as outputs are FITS, FTCL parameter files, CSV files for database loading, and XML files. For each set of data that is produced (raw data, meta-data, reduced data, integration data) there will be a "data model": HTML files that describe for each type of file the directory location, naming convention, required and optional header keywords, required and optional column names, and any relevant data types, units, and descriptions for each. The data model documentation is stored in an SVN product ("sas") and a periodically updated version distributed on the web site http://sdss3.org/internal/datamodel. Changes to the data model must be approved by the Data Coordinator.

The total archival needs for SDSS-III are estimated to be 100 Tbytes. About 50% is raw data and 50% reduced. About 70 Tbytes are required by July 2009, and the needs grow by 6 Tbytes/year thereafter. We assume here that APOGEE does not require Fowler sampling, which would increase the raw data storage needs. The full breakdown is shown in Table 13.1 below. Final calibrated catalog information (about 15 Tbytes) should be on robust servers with 100% up-time; raw data and spectra will be on commodity hardware (with multiple backups to handle failures smoothly).

SDSS-III Data Volume	July 1 2008	July 1 2009	July 1 2010	July 1 2011	July 1 2012	July 1 2013	July 1 2014	
raw								
SDSS-I, -II spec	2	0	0	0	0	0	0	**2**
SEGUE-2	1	0	0	0	0	0	0	**1**
SEGUE-bright	0	0.5	0.5	0.5	0.5	0.5	0	**2.5**
BOSS spec	0	1	1	1	1	1	0	**5**
MARVELS	1	1	1	1	1	1	0	**6**
SDSS-I, -II imaging	28	0	0	0	0	0	0	**28**
BOSS imaging	2	0	0	0	0	0	0	**2**
APOGEE	0	0	1	1	1	1	1	**5**
reduced								
SDSS-I, -II spec	1	0	0	0	0	0	0	**1**
SEGUE-2	1	0	0	0	0	0	0	**1**
SEGUE-bright	0	0.5	0.5	0.5	0.5	0.5	0.5	**3**
BOSS spec	0	0.5	0.5	0.5	0.5	0.5	0.5	**3**
MARVELS	0.5	0.5	1	1	1	1	1	**6**
SDSS-I, -II image cats	14	0	0	0	0	0	0	**14**
BOSS imaging cats	2	0	0	0	0	0	0	**2**
BOSS datasweeps	1	0	0	0	0	0	0	**1**
Imaging frames	0	10	0	0	0	0	0	**10**
APOGEE	0	0	0.5	0.5	0.5	0.5	0.5	**2.5**
Data integration	2	1	0.5	0.5	0.5	0.5	0.5	**5.5**
increment	55.5	15	6.5	6.5	6.5	6.5	4	
total	55.5	70.5	77	83.5	90	96.5	100.5	**100.5**
increment (stable)	5	2	1.5	1.5	1.5	1.5	1.5	**14.5**
total (stable)	5	7	8.5	10	11.5	13	14.5	14.5

Table **13.1**: SDSS-III data volume needs, broken down by survey and data type. Each year's increment is listed. Some data we need to keep on very robust servers (rather than commodity hardware). Out of the full 100 Tbyte, these data account for about 15 Tbyte, which we track in the final rows.

13.2. Data Processing

Data processing activities include all work associated with the development and maintenance of the data processing pipelines, and the organization, operation, and maintenance of the data processing factories.

As described in Section 1.8, SDSS-III data processing is divided among several institutions: LBL for BOSS imaging and spectroscopy; Princeton for SEGUE-2 spectroscopy; University of Florida for MARVELS spectroscopy; and University of Virginia for APOGEE spectroscopy. The PI for each team is responsible for each data processing unit. On-mountain processing occurs at APO for quality assurance purposes. The Lead Observer is responsible for the on-mountain processing. Data assembly steps such as window function determination and matching of imaging to spectra occur at SAS. The Data Coordinator is responsible for the data assembly and for setting the overall schedule for reduction completion.

BOSS imaging reductions will produce a set of photometric catalogs, astrometric and photometric calibrations, datasweep files, and targeting and tiling results.

BOSS spectroscopic reductions will produce calibrated spectra for each fiber, calibration results for each fiber, spectroscopic parameters associated with these spectra and quality flags on each spectrum.

SEGUE-2 spectroscopic reductions will produced calibrated spectra for each fiber, calibration results for each fiber and spectroscopic parameters associated with each spectrum. The SEGUE-2 team will also produce the targeting information associated with each field.

MARVELS spectroscopic reductions will produce calibrated spectra and spectroscopic parameters for pre-selection plates observed with the SDSS or BOSS spectrographs, and extracted results for each fiber along with radial velocities from the MARVELS spectrograph. The MARVELS team will also produce the targeting information associated with each field.

APOGEE spectroscopic reductions will produce wavelength-calibrated, telluric-absorption corrected spectra for each observation, co-added spectra for multiple observations of the same star,, and associated spectroscopic parameters (radial velocity, atmospheric parameters, abundances, and uncertainties). The APOGEE team will also produce the targeting information associated with each field.

For each pipeline that is run, the outputs will track the version of the pipeline and the reprocessing number (as we describe in Section 15.3, the versioning and dependency trees are tracked with SVN and ExtUPS).

The data assembly step is special in the sense that it incorporates multiple survey inputs and produces science-ready files, with associated statistical weight maps and random points where appropriate. The Data Coordinator is responsible for this processing step, relying on help as appropriate from the science teams. This assembly also includes matching objects across the various surveys.

Data assembly depends on input from external data sets, in particular the SDSS-I and SDSS-II surveys and geometry files. The Data Archive Scientist will be responsible for maintaining all external data sets except for the SDSS-I and -II legacy data, which are the responsibility of the Data Coordinator.

Where surveys depend on reductions of other surveys' data (such as APOGEE on SEGUE-2, or BOSS spectroscopy on BOSS imaging, or MARVELS on pre-selection plates from SDSS) they will copy the necessary reductions from SAS.

Data reduction quality is checked through a series of checks described in Chapter 14.

13.3. Data Distribution

The data distribution for SDSS-III is performed through the Science Archive Server at LBL and the Catalog Archive Server at JHU. Mirrors of the CAS will be run at several international locations (Europe, Asia, and South America) to provide fast local access world-wide. A Science Archive Mirror exists at NYU to mirror the LBL site, but it is not expected to be widely used for distribution.

The SAS is run at LBL and maintained by the Scientific Cluster Support team there. It will allow access through an HTTP interface (e.g. available to web browsers, wget and curl) as well as read-only rsync support. A web front-end will allow simple search and retrieval operations, in particular for images and spectra. Private data will be available to collaborators through a password-protected interface. Public data will be available without password protection. The Data Manager will oversee the development of this web site.

The SAS will be mirrored offsite at the SAM (at NYU). The Data Archive Scientist will maintain this mirror.

The CAS is run at JHU and is maintained by the Database Development Group there. It allows access through a MicroSoft SQLServer interface that allows for very flexible querying. An associated CASJobs interface allows server-side caching of and operations on intermediate results. A lay-level front-end called SkyServer distributes navigable JPEG images of the sky, with labeling referring back to the full database and GIF images of spectra for visual inspection. Private data will be available to collaborators through a password-protected interface. Public data will be available without password protection. The CAS Head oversees the development of this database. One dedicated FTE (possibly split across two individuals) will administer the database.

The CAS will be mirrored at multiple other sites, to be determined. The main database, but not CASJobs or SkyServer, will be required to be included in the mirror. Contacts at those sites will administer those mirrors, and coordinate update with the CAS Head. Mirror sites may run CASJobs are SkyServer if desired, but ARC does not take responsibility for maintaining those sites.

Loading of the database will occur in a staged manner. Calibrated FITS files will be created on SAS that are as nearly parallel as possible with the final CAS tables. These FITS files will be converted to temporary CSV files that will be transferred to JHU (either on portable disk storage or over the network) for loading into the database. CAS requires some internal calculations to occur before it is ready for consumption.

As close to 100% uptime as possible will be maintained on SDSS-III CAS and SAS servers serving the current data release to the public.

We have created a testbed version of the SAS, SAM, and CAS systems, with a small subset of the plates and imaging runs from the full survey. For testing development and changes in procedures, we use the testbed system.

13.4. Data Documentation

The details of the data in SAS and CAS will be documented. The Data Coordinator is responsible for assembling this documentation, with the assistance of the survey teams as necessary.

To describe the data at SAS, a data model is maintained in the “sas” SVN product and posted at http://sdss3.org/internal/datamodel. This data model contains a description of each archived file type and its required contents (and any known optional content), with the location of the file in the global directory structure, plus the meaning, data type and units of each entry. The data web server will have description of how to extract the data as well as examples of how to do so.

The algorithms used for producing reduced data will be documented on the http://sdss3.org site (as well as in technical papers of course).

To describe the data in CAS, each SQL table has descriptions, units, and data types for each item. In addition an algorithm page describes algorithms used in CAS itself. The CAS has documentation describing its use and as well as examples of how to do so.

In addition to providing online documentation, Jordan Raddick (JHU) will continue to offer programs to train the astronomical community in using the SAS, the CAS, and CasJobs to retrieve SDSS-III data. These programs will be based on the "Cooking with Sloan" trainings offered at previous AAS meetings, and the online follow-up materials for Cooking with Sloan will be adapted for SDSS-III. Training programs will be given twice a year as Special Sessions or Splinter Meetings at AAS meetings; there is also a small travel budget to pay for trainings at other astronomy meetings or invited trainings at university astronomy departments.

13.5. Data Releases

As observations progress, data archived on the SAS will be available to SDSS-III participants as it arrives, through the password-protected web server.

The CAS databases will be loaded periodically. Typically, we aim for the databases to be fully loaded six months or more previous to each public data release, to provide lead time for participants as well as an opportunity to vet the data. The Director and Project Scientist must approve the release of the data to the public.

The data release schedule is below. In most cases, we are aiming for a data release within a year after each observing season is completed. We include a DR11 “internal” release that includes data from the 2012-2013 season; with the full release scheduled for Dec. 2014, a public release the previous summer is unwarrant

Date	Data Release	APOGEE	BOSS	MARVELS[a]	SEGUE-2
Dec. 2010	**DR 8**	---	**Imaging** (up to Jan. 2010)	---	**Spectra** (up to July 2009)
July 2012	**DR 9**	---	**Spectra**[b] (up to July 2011)	**Radial velocities** (up to July 2011)	---
July 2013	**DR 10**	**Spectra**[c] (up to July 2012)	**Spectra** (up to July 2012)	---	---
July 2014	DR 11 [internal][d]	Spectra[c] (up to July 2013)	Spectra (up to July 2013)	---	---
Dec. 2014	**DR 12**	**Spectra** (complete)	**Spectra** (complete)	**Radial velocities** (complete)	---

Notes: (a) MARVELS will only execute a public release if additional funding is identified. (b) Initial BOSS spectroscopic reductions, released in DR9, will use the preliminary BOSS pipelines. (c) APOGEE intermediate data releases will only include completed integrations of stars (not partial integrations). (d) The internal data release DR11 is to help iron out problems and prepare the collaboration for the final release a few months later (this approach was invaluable for SDSS-II).

13.6. Data Distribution Reviews and Milestones

Here we lay out the various milestones for the data distribution effort, along with their planned completion time. The database loads are associated with acceptance reviews of data quality assurance tests and functionality. The testbed database is a small fraction of the total data that we use for experimentation and testing.

Archiving system for raw data developed: Summer 2008 (completed)

Archiving system for reductions developed: Fall 2008 (completed)

SAS web service available to collaboration: Spring 2009

Large-scale structure samples developed: Summer 2009

CAS database available to collaboration: Fall 2009

CAS mirroring tools completed: Spring 2010

Testbed database loaded [Acceptance review]: Summer 2009

DR8 database loaded [Acceptance review]: Winter 2010

DR9 database loaded [Acceptance review]: Fall 2011

DR10 database loaded [Acceptance review]: Winter 2013

DR11 internal release database loaded [Acceptance review]: Winter 2014

DR12 database loaded [Acceptance review]: Fall 2014