LSST DESC Notes

Summary of Data Management Principles

The LSST Dark Energy Science Collaboration (DESC)
Last revised: January 9, 2020
Version 1

Experiment Description

The Legacy Survey of Space and Time (LSST) Dark Energy Science Collaboration (DESC) was established in June 2012 with the goal of developing and executing a high-level plan for the study of dark energy and associated fundamental physics of the Universe with LSST data. DESC is one of several science collaborations that is preparing for scientific analysis of LSST data. The LSST Project has its own Summary of Data Management Principles; this document covers only LSST DESC-produced data products, though LSST will be described briefly in order to motivate the DESC data management principles.

LSST: LSST will be carried out at the NSF Vera C. Rubin Observatory (VRO) with the Simonyi Survey Telescope, which has an effective 6.7-m diameter primary mirror, and the DOE LSST Camera, which has a 9.6 square-degree, 3.2 Gigapixel camera and is equipped with 6 optical filters covering the wavelength range 320 – 1050 nm. Over 10 years of operation, LSST will perform a minimum of 825 visits of every part of the southern sky. The VRO is presently under construction on Cerro Pachon in central Chile, with the official start of survey operations in late 2022. A data management system is under construction to retrieve, process, analyze, and archive the massive data volume, approaching several hundred Petabytes. Users of the data (including the LSST DESC) will initially access the data through LSST Data Access Centers (DACs), which will enable database queries, a compute-limited amount of scientific analysis, and bulk downloads of the data.

DESC: DESC is preparing for cosmological analysis of the LSST data; science requirements are driven by the goal of understanding the accelerated expansion rate of the Universe, which is attributed to the poorly understood dark energy. Work within DESC is centered on five primary probes of dark energy: weak and strong gravitational lensing, large-scale structure, galaxy clusters, and supernovae. DESC is not directly involved in the operation of the LSST hardware or basic data management system; these will be run by LSST Facility+Operations. DESC's efforts are instead focused on developing analysis methodology and software infrastructure to

support the aforementioned dark energy analyses, and providing feedback to LSST about the impact of survey strategy, image processing, etc. on dark energy science. DESC is an international collaboration and is thus distributed geographically. DOE's National Energy Research Scientific Computing Center (NERSC), operated by LBNL, is the primary DESC resource for data reprocessing, analysis, and general data access for DESC members. DESC has previously secured large storage resources and is now using the new NERSC Community Storage to enable data serving and archiving at NERSC. In addition, currently three major secondary computing resources exist at international partner institutions (CC-IN2P3 and in the UK) and a DOE computing facility - the Argonne Leadership Computing Facility (ALCF). These resources are being used for simulations, data processing, and analysis. DESC may partner with an LSST alert broker to facilitate aspects of its time domain science cases.

The DESC has a Science Roadmap (SRM) that outlines the work the collaboration is undertaking to prepare for a robust and timely cosmological analysis of LSST data, including the development of simulation, reprocessing, and analysis software that works at the necessary scale and precision for LSST. Of relevance to this document is the fact that DESC will generate its own data products, specifically (a) *simulated datasets* of varying complexity to enable the development and validation of analysis software, (b) the outputs of *reprocessing (subsets of) the LSST data* to understand systematic uncertainties, and (c) *value-added data products* based on the LSST data, such as catalogs of galaxy clusters.

DOE's roles in the experiment

The Department of Energy Office of High Energy Physics considers LSST to be a Stage IV Dark Energy Experiment. It is supporting LSST DESC Operations, in addition to the development and fabrication of the LSST Camera, and ~50% of the cost of operating the full facility. The DESC computing model is currently built around support from DOE computing facilities augmented by contributions from international partners.

Partnerships

The US DOE funds DESC Operations, which also benefits from in-kind contributions from institutes and DOE labs within the US, and from international partners CC-IN2P3 (France) and STFC (UK), which play key roles in DESC computing as mentioned above. Communication between the DOE and international funding agencies is organized through an International Resource Committee.

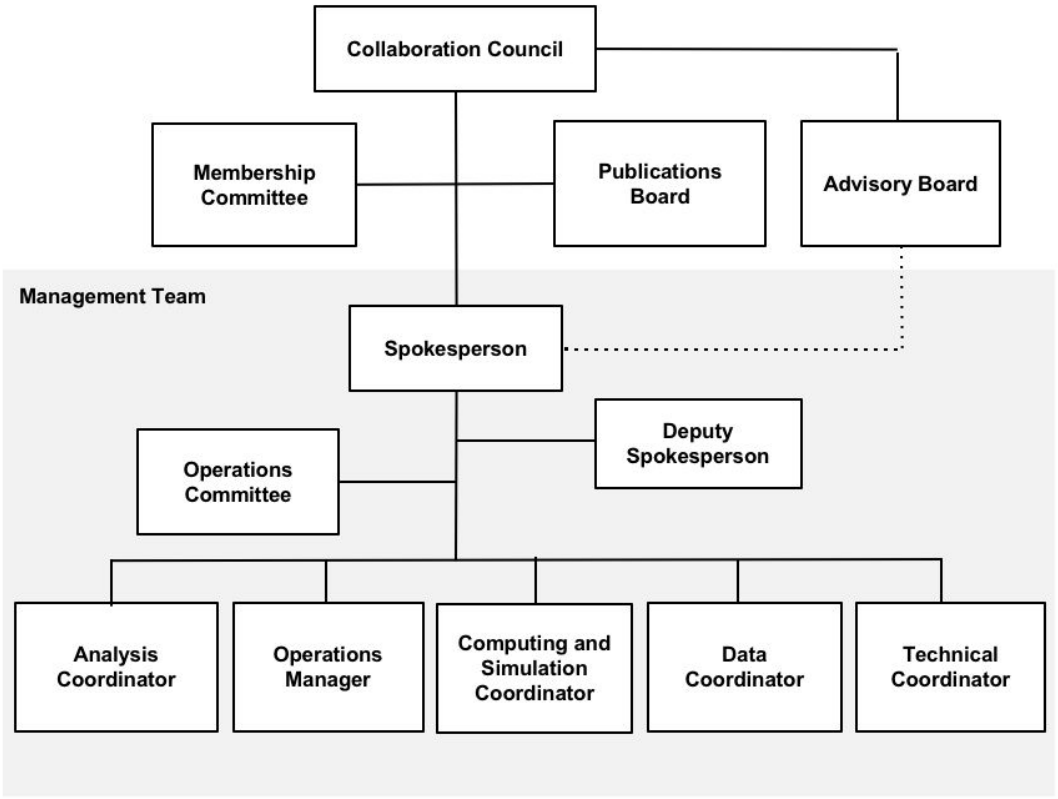
Since the LSST DESC is an international collaboration, the scientific activities of its members are funded by a variety of agencies, including both the US DOE and NSF. DESC has also benefited from support from the LSST Corporation Enabling Science program.

Organization - Agency/Lab Level

SLAC is the host laboratory for the LSST Dark Energy Science Collaboration.

Organization - Experiment Level

The chart below shows the internal organization of the LSST DESC. Connections to the host lab are facilitated by direct engagement of the DESC Spokesperson with a specified point of contact at SLAC, and through the Operations Manager, who is based at SLAC.



Collaboration

The DESC was established in June 2012. As of December 2019, it has ~220 full members (with a high level of commitment to the collaboration and a role in its governance) and ~1000 members total, of whom ~70% are based in the US and the rest are distributed across ~15 other countries, with large concentrations in the UK and France. The DESC has an elected Spokesperson who serves for a two-year term. The Spokesperson appoints the other management and leadership positions (those that report to the management team) within the

collaboration, with the Operations Manager being appointed in consultation with SLAC; management roles require confirmation by the Collaboration Council. The Collaboration Council is elected by the Full Members of DESC, and establishes the membership of the three committees shown directly below it on the org chart. DESC membership is restricted to those with LSST data rights.

Data Policy Management

The DESC's data management plan is currently under development and is being overseen by the DESC Management Team. Key aspects of how we will manage and serve data are already known and are described below. Detailed responsibilities for implementing the data management plan are still being determined while it is under development; oversight in its implementation will be provided by the DESC Management Team. Resource allocation according to DESC priorities, including resources needed for carrying out this plan, falls under the purview of the DESC Operations Committee.

Data Description and Processing

Simulated data: DESC produces simulated LSST datasets of varying complexity to enable the development and validation of analysis software. The simulated data is processed by the LSST Science Pipelines, and made available to DESC members at NERSC. Image simulation and processing are carried out across the primary and secondary DESC computing resources.

LSST data: LSST Operations is responsible for running the LSST Science Pipelines and serving the imaging and catalog data and tools for interacting with them to the LSST data rights community, including DESC members, through LSST Data Releases. DESC will interact with LSST Data Releases through the US LSST Data Facility for initial validation tests, but for full validation and resource-intensive analysis tasks, the intention is to bulk-download the catalog data and selected subsets (<~10%) of the imaging data to NERSC; this must be arranged through an agreement with LSST Operations as per the LSST data policy, http://ls.st/LDO-13 (DPOL-511).

DESC primary data products will be the supporting catalogs for each of the five main cosmology probes. These will be the data products of most direct use by the community beyond DESC. The primary data products will include photometric redshift and classification, weak lensing shear maps, cluster counts, and light-curves of variable quasars and supernovae. These data will be primarily catalog data, supplemented by pixel data for the transient and variable events, and completeness masks for the correlation functions.

DESC secondary data will include re-processed subsets of the full LSST imaging dataset to validate analyses and to explore systematic effects. These datasets will be large (~2 PB persisted, with 10-20 PB needed during reprocessing) with multiple re-processings of individual images and coadded stacked images. These will be used to generate systematic error budgets

and checks on the quality of the delivered data. Most internal (DESC) and external users will be interested in the summarized results of this processing, but not the intermediate results. Processing is planned to be undertaken at NERSC, with key additional capabilities provided by US and international partners.

Data Products and Releases

Simulated data: The timing at which we will make simulated data releases available beyond DESC has not yet been defined; however, DESC is unlikely to have the resources to serve the full simulated imaging datasets for our data challenges. Releases of coadded images and object catalogs are planned, though the timing of these releases is still to be determined when finalizing a data management plan.

LSST data: We anticipate that DESC Key Papers will be based on LSST Data Releases. With those Key Papers, DESC will release data products that build on the LSST data products and feed into DESC cosmology analyses, potentially including additional, non-LSST data (e.g., from spectroscopic follow-up). The DESC data products that are provided will be clearly linked to the LSST Data Release on which they are based and will be released in support of the relevant DESC Key Papers. The plans for releases of DESC data products based on LSST data will respect the LSST data policy constraints on when data products can be released, which depends on whether they are categorized as Derived Data Products (c.f. section 6 of the LSST data policy, http://ls.st/LDO-13).

Plan for Serving Data to the Collaboration and Community

Both simulated and real LSST data products produced by DESC will be served to DESC members through NERSC. We may provide the data through additional international and US DOE resources to enable expanded analysis capabilities and flexibility.

Relevant data products produced by DESC that support Key Papers will be made available beyond DESC in conjunction with the publication of the associated papers, in a tiered fashion described below.

Plan for Archiving Data

DESC-produced simulated and derived LSST data products used for publications are planned to be archived at the NERSC High Performance Storage System during the LSST and for at least 10 years afterward subject to DOE funding availability.

The code and environment used to produce an analysis will be archived as full images, such as Docker containers, sufficient to recreate the environment in which the code was executed.

Plan for Making Data Used in Publications Available

Data products and software sufficient for reproducing the top-level results in DESC Key Papers will be made available, with the timing to be established in the DESC data management plan (under development). Both simulated data products and those based on real LSST data will be provided for download through NERSC. Software will be made public on GitHub. We will additionally provide the Docker-like containers used to execute the code.

The data will be provided in a tiered manner. Most immediately, at the time of publication, we will provide the data points in the published plots in machine-readable form. We will then provide increasingly detailed (and likely larger) data files as we step backward along the analysis chain.

The DESC infrastructure will support making data products and code public for DESC non-Key Papers.

At the time of publication reproducing the full analysis chain may require being an LSST Data Rights holder, but once the data from the LSST Data Release used in the DESC are non-proprietary, reproduction should be possible by anyone with sufficient computing resources.

Responsiveness to the DOE Office of Science Statement on Digital Data Management

This data management plan fully follows the Office of Science Statement on Digital Data management:
https://science.osti.gov/Funding-Opportunities/Digital-Data-Management