

DES Data Management Formal Production Organization

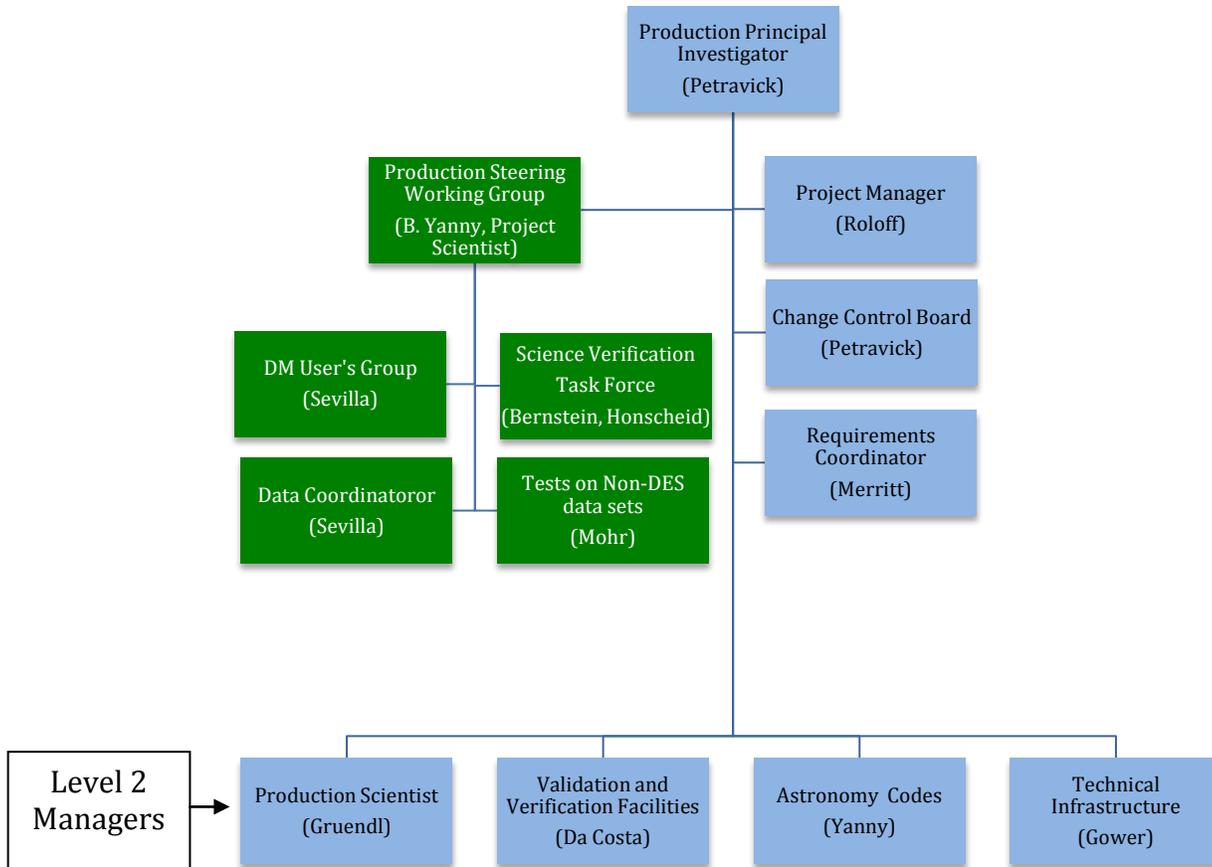
D. Petravick
Version 6.9, Aug. 1, 2012
This is DES Doc-Db-6594

1 Introduction

This document outlines an organization for the DES Data Management Formal Production system. This system is a component of the overall organization of the Dark Energy Survey. The system archives, processes, and serves the survey's core data sets. Some of these data sets are produced within the DES production system, others are value-added data products produced by collaboration members, but accepted for incorporation into the survey data sets.

The organization provides for the continuing development of both the science codes and of the data management system itself. An appendix gives the organization for the community pipeline, which is derived from the DES systems, but unlike DES Data Management, is not funded by the NSF.

The organization chart is effective once approved by the DES Director.



2 Production Principal Investigator

The Production Principal Investigator (PI) organizes the overall production effort. The PI maintains a close relationship with the DES Project Office. The PI has a line relationship with the four level 2 managers, and is advised by the Production Steering group, chaired by the DESDM Project Scientist. The PI is assisted by the Project Manager and the Requirements Coordinator.

The PI further:

- Chairs the DESDM Change Control Board.
- Is advised by the Production Steering group.
- Holds production priority decision authority.
- Holds overall accountability for the organization.
- Allocates DESDM and CP funding
- Coordinates institutional contributions.

2.1 Change Control Board

The change control board manages the approval and sequencing of changes (both science and non-science) into the formal project plan.

- Organized and Chaired by Production PI
 - Members: Level 2 managers, Project Scientist, Project Manager, Requirements Coordinator
- Approves flow-down of other-than-science requirements for system (Project Scientist is responsible for science requirements).
- Responds to science requirements.
- Oversees execution of changes to the system.
 - Maintains prioritized list of needed changes
 - Plans and sequences changes to the system.
 - Evaluates the success of changes to the system

2.2 Project Manager

- Assists the production PI with the planning and execution of the project plan
- Manages the project budget and schedule

2.3 Requirements Coordinator

- Maintains comprehensive set of leveled DESDM requirements.
- Assesses completeness of implementation.
- Assesses completeness of testing of DESDM requirements.

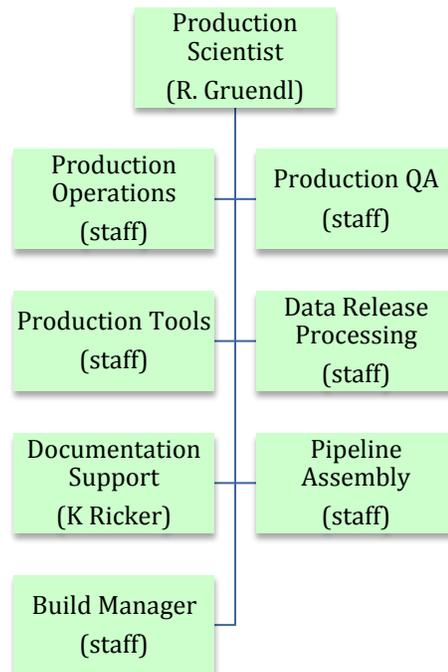
2.4 Project Scientist

- States Scientific Priorities.
- Chairs the Production Steering Working Group (below)
- Approves flow-down of DES science requirements to DESDM requirements through the requirements coordinator.
- Coordinates/ensures delivery of Astronomy Codes (see below)

2.5 Production Steering WG

- Organized, Chaired, by the Project Scientist
- Receives input from
 - Level 2 managers
 - Coordinates activities that are not part of production
 - DM User's group
 - Tests on non-DES data.
 - Science Working Groups input to DESDM
 - Science Verification Task force feedback to DESDM
- Receives QA data from the production system.
- Makes consensus statements about:
 - Priorities for evolving the system.
 - Priorities for sequencing improvements into the system.
 - Production priorities.

3 Production Scientist



- Production Scientist: Responsible for overall production organization.
- Production includes:
 - Runs to support software and systems development.
 - Runs to qualify software for production.
 - Runs to study alternate calibrations, science configurations, small software changes.
 - Runs to produce survey data products.
 - Methods for uniform operations under controlled conditions.
 - Initial assessment of the data.
 - Assembly of functional units into pipelines.
- Responsible for methods of production, e.g. amount of QA within DESDM.
- Sits on Production Steering Working Group .
- Responds to production priorities set by the Production Steering WG.

Production Operations

- Runs pipelines for larger scale tests
 - benefitting software development,
 - fit-for-use testing,
 - calibration and science parameter studies
 - formal production.
- Retries to mitigate technical failures,
- Retry with alternate software version
- Retry with alternate calibrations or configurations
- During observing season,

- focuses on daily cadence for arrival driven processing.
- flags data not passing its level of controls for production QA.
- Passes data to Production QA for secondary assessment
- Reports progress and records discretionary action and initial assessment in logbook and other data artifacts.

3.1 Production QA

- Reviews output from Production Operations.
- “blesses” data for retention, or marks as defective.
- Analyses problem data to level of credible bug report.
- Maintains list of open problems.
- Maintains impression of the operational severity of each problem.
- Dispatches selected data for science WG validation.
- Fills out survey table.

3.2 Production Tools

- Provides survey-specific tools to mark data.
- Provides Python-based “offline quick look” tool base which is aware of the survey data model.
- Provides specific offline evaluation tools.
- Provides Quality controls embedded in pipelines filling any gaps left by science codes.
- Provides cutout server
- Provides “footprint server”

3.3 Data Release Processing

- Transforms, validates catalog and file data into release formats
- Generates release notes and other production generated release documentation.
- Generate model queries and other technical materials.
- Announces release.
- Incorporate value added data products

3.4 Documentation Support

- Provide specific methods for data documentation built on
 - Confluence
 - one-line summary available in database.
- Review documentation provided by science code projects.
- Deals with documentation bugs.

3.5 Pipeline Assembly

Assembles units of functionality into test and production pipelines e.g.:

- Precal
- SuperCal
- FirstCut
- FinalCut
- SNE SE

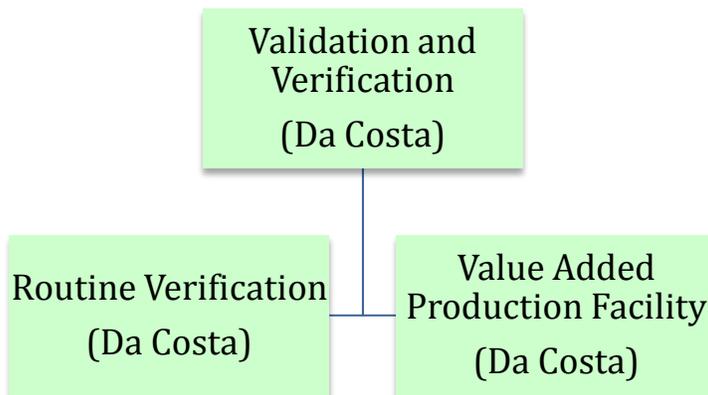
- Diff Imaging
- Coadd
- SE WL
- ME WL
- Local and Global calibration

3.6 Build Manager

The primary function of the build manager is to incorporate new software and configurations which have been unit-tested as part of their delivery from the Astronomy Codes group into a branch that is targeted for a release, and validate them over nominal, simulated and corner-case test data.

The build manager ensures that accepted changes for a specific release are merged back into the main body of software. Priority is given to features sequenced into the system by the change control process.

4 Validation and Verification



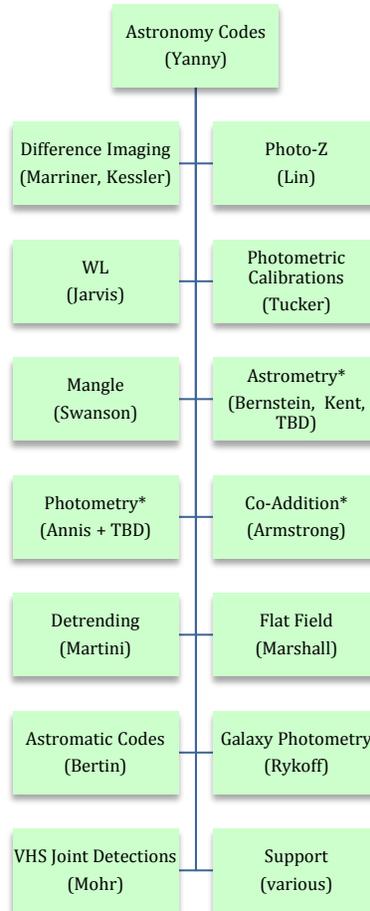
4.1 Routine Verification

- Provide an infrastructure for independent verification (I.V.) of DESDM data sets.
 - Receives release and pre-release data make available for I.V.
 - Provides framework for hosting codes for I.V. along with software and configuration management.
 - Identifies a process for accepting codes identified as verification codes.
 - Provides for routine running of I.V. codes at the facility.

4.2 Value added Dataset Production Support.

- Hosts science analysis. Part of this is providing a mechanism for making value-added data sets generated at the analysis facility ready for ingest into the main

survey body of data, Criteria are similar to the science topic data under the project scientist.



5 Astronomy Codes

- Coordinates development of science codes.
- Oversees algorithmic soundness of scientific codes.
- Oversees interoperation and data interfaces between codes.
- Coordinates changes, and is responsive to change priorities.
- Coordinates with Production Scientist in incorporating/integrating science codes into the production system.
- Sits on Production Steering Group.
- Works with the DES project office to staff science code efforts.
 - Each “box” might require several skills.

5.1 Science Topics

Science code organizational units presented in the organization chart provide the following aspects of the code for the topic area, conforming to survey methods and standards, including:

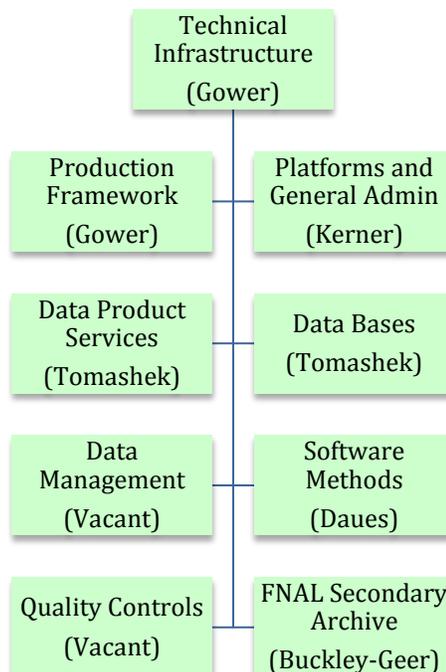
- Code itself, licensed for use by the survey.
- Essential quality controls exposed.
- Unit level test and testing.
- Documentation of data outputs.
- Science level configuration files, or other configuration data.
- “Wrapper” adapting the software to the production framework.
- Support analogous bits of the CP, if applicable, for the duration of the support CP obligation.
- respond to science code change priorities

As the system matures, we anticipate fewer organizational sub-units.

5.2 Support

There is a tension between finding competent scientist to oversee this work and the technical deliverables needed for each science topic. The support function provides a limited amount of assistance to the scientist responsible for each science area. The support effort is limited and is mostly provided by institutional contribution (throughout the collaboration).

6 Technical Infrastructure



6.1 Technical Infrastructure, Lead.

- Oversees DESDM's Technical Base. Sits on steering WG.
- Provides the technical infrastructure.
- Responds to technical change priorities.
- Long-term issue mitigation with high-performance-computing resource providers.

6.2 Production Framework

Supplies and maintain software and services needed for batch processing,

Submit a run

Monitor runs submitted/running

Run restart (not focus of development, but trying to avoid causing problems if we will need restarts, currently at block level - failed setup, failed file ingest, OP manually finished block start next block)

Get historical run information

DB storage of PFW meta-data (run, block, start times, end times, etc)

Interface with Condor

Monitoring: Operators interactive + push notifications (email, jira, wiki)

WCL reader/writer)

Querying database for initial inputs

Dispatch work into workflow units.

Creating WCL inputs for wrappers (list input files, required metadata, config info, etc)

Creating runtime python workflow for inside each target job (

Register output files with DB

metadata,

parentage

software provenance

Community code provenance

Provide/Integrate QCF (ability to turn on or off)

6.3 Platforms and General Administration:

- Construction/provisioning/operations/support/ architecture of DESDM cluster (floor 3 NSCA) includes central file store.
- Account provisioning for all account and services., including authentication and authorization for all systems, including
 - o UNIX/File systems
 - o Services such as confluence, etc.
- Provide installation, maintenance, licensing of third-party service packages from commercial or community or NCSA suppliers, including.
 - o Jira,
 - o Confluence,
 - o svn,
 - o Gatekeeper,
 - o Globus,

- VAO
- Apache
- Condor
- And similar.
- Provide computer security planning and execution consistent with NCSA policy.

6.4 Data Product Services

Provides external data interfaces to collaboration and public for the use cases of

- Person with data rights
- Collaborators accessing non release data for purposes of advancing the survey.

Provides

- supported tool for database access (E.g. trivialAccess).
- File pick (via HTTP)
- Bulk file downloads FTP/GridFTPVAO services

6.5 Databases

- Provide Oracle RAC and other database hardware, hardware support, and other infrastructure need to maintain relational databases.
- Provide Oracle instances needed for production and test.
- Provides operational support, such as backup/recovery, disaster recovery, monitoring tuning, and database specific operational effort.
- Provides database tables, schema, and other design artifacts, items including
 - Provenance and meta-data schema.
 - File catalog schema.

6.6 Data Management

- Provides a file storage architecture, catalog and other software items needed to maintain the DESDM file base.
- Provides file management software that is sensitive to provenance and other information relevant to file management.
- Provides Disaster recovery for the file base.
- Provides for file management, and movement within the production system.
- Provides methods for input and output of files from the system
- The liaison with and ingest tools for the NOAO Data Transport System.

6.7 Software Methods and Support

- Provides standards and methods for software development and small scale testing prior to larger scale testing by Production Operations.
- Maintains a package management system
- Maintains a repository of versioned software which may be deployed systematically and efficiently to the remote computing resources, and maintain the capability to advance or backtrack the software between versions as needed for stable production processing.
- Manages and maintains any software licenses

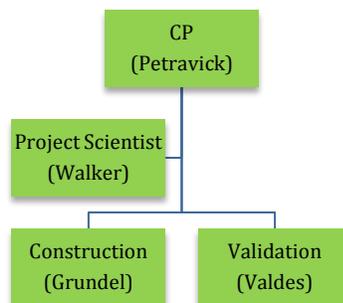
6.8 Quality Controls

- Provides infrastructure to extract and store QC data from varying sources like log files, (run time notifications?) and databases.
- Provides tools and methodology to mine the QC data for **predictive and diagnostic analysis** of the health of the processing system
- Enables a layer of intelligence around thousands of stored log files for effective information retrieval
- Provides actionable information to operators during a processing run

6.9 FNAL Secondary Archive

- This is a copy of the DESDM science archive.
- In the event of a problem at NCSA, the Fermilab site serves as a backup, enabling access to DES data releases by the Collaboration until the problem is resolved.
- When not serving in the backup role, the Fermilab mirror site handles queries put to the database by the Collaboration, thus supplementing computing resources at the primary archive.
- Finally, the secondary archive is an essential part of the Fermilab astrophysics plan for exploiting DES data for science analysis.

7 Appendix: Community Pipeline Organization Chart



The Community Pipeline is a deliverable from the DES collaboration to NOAO and forms part of the DECam System. It is derived from DESDM and is being constructed under the leadership of the DESDM team at NCSA. The pipeline is separately funded by the DES collaboration. (NSF funds may not be used for its construction and support). It will be operated by NOAO for the astronomical community, to process non-DES data from DECam. The relation of the CP team to the DECam Operations organization led by NOAO is described in the DECam System Operations and Maintenance Plan.

The Project Manager, Project Scientist, Construction Manager, and Validation Manager meet weekly in a status meeting that is also the Change Control Board.

The construction project maintains a monthly cadence of releases to NOAO. This process requires the following skills:

- A senior astronomer to liaise with NOAO, and direct the overall processes, and liaise with the programmers as features are developed.
- Programmers to implement changes. Two skills are needed: Science coding is typically in the “C” programming language, and “framework” changes in perl.
- A tester – who integrates into the pipeline and tests at scale, works bugs with the developers.
- A release team which controls the code in the week prior to release, validates, writes release notes, makes a distribution available to NOAO, and announces the release.
- Once CP development is completed and CP operations on-going at NOAO, DES/NCSA will be responsible for bug fixes to the DES-supplied code, as described in the DECam System Operations and Maintenance Plan.