DataLab-OCD-v0.90

Operational Concepts Document

For the

NOAO Data Lab Project

Revised: March 3, 2015

Revision History

Date	Author	Changes / Comments	Version
Aug 14, 2014	M. Fitzpatrick	First Draft	0.1
Aug 30, 2014	K. Olsen	Merged input	0.2
Sep 22, 2014	M. Fitzpatrick	Additional text	0.3
Oct 07, 2014	M . Fitzpatrick	Added concept ids	0.4
Oct 17, 2014	M. Fitzpatrick	Added tech overview text	0.5
Oct 31, 2014	M. Fitzpatrick	Concision revision	0.6
Nov 11, 2014	M. Fitzpatrick	Restructuring	0.7
Nov 14, 2014	M. Fitzpatrick	Add'l text	0.72
Nov 14. 2014	M. Fitzpatrick	Complete first draft	0.8
Nov 19, 2014	M. Fitzpatrick	Formatting / Tracked ORD	0.82
Nov 24, 2014	M. Fitzpatrick	ORD revisions	0.83
Dec 04, 2014	M. Fitzpatrick	Ridgway comments	0.84
Dec 04, 2014	K. Mighell	Added SRD tags	0.85
Jan 23, 2015	K. Mighell & B. Stobie	Added new acronym list	0.86
Mar 03, 2015	K. Mighell	Edits	0.90

Table of Contents

1 Document Overview

1.1 Purpose

This *Operational Concepts Document* for the NOAO Data Lab has two purposes:

a) To describe and define the components of the Data Lab in a way that the operational requirements can be derived. This complements the science requirements that specify *what* the Data Lab should accomplish by describing *how* it will be done.

b) To provide a high-level overview of the system from the perspective of the scientists and developers who will use it by describing how they will interact with the various components.

This document first describes the science capabilities of the Data Lab identified in the Science Requirements document. It then describes the infrastructure components to be built and their operational functionality.

The scope of this document is the entire Data Lab Project. This document may evolve over time as requirements and designs are finalized.

1.2 Referenced Documents

This document may reference additional documentation identified below.

[1] Science Use Cases	(SUC)
[2] Science Requirements Document	(SRD)
[3] Operational Requirements Document	(ORD)
[4] System Architecture Design	(SAD)
[5] Program Execution Plan	(PEP)

2 Overview and Scope of the Data Lab

2.1 Introduction

NOAO telescopes have been deeply involved in survey science for more than a decade. In particular, the NOAO Survey Program, which was established in 1999, has provided some of the highest impact science obtained on NOAO facilities. This trend is bound to continue with the Dark Energy Camera (DECam), the Dark Energy Survey (DES), and the Dark Energy Spectroscopic Initiative (DESI) leading the community to the era of LSST.

Despite the overall success of NOAO survey science, its potential impact could be significantly greater. Two current weaknesses are (1) the limited interface to access survey data that the NOAO

Science Archive provides and (2) the inability to publish, distribute, and provide access to derived survey data products, in particular catalogs, through NOAO. With DECam and DES already upon us, and the approaching horizons of DESI and LSST, the community has a big opportunity to maximize the impact of survey science through NOAO.

The NOAO Data Laboratory is intended to help the community take advantage of current large surveys and to prepare them for the era of LSST. The Data Lab will allow users to:

1. Access, search, and filter databases containing large catalogs
2. Create custom databases and analyses from large catalogs using familiar tools
3. Combine catalog databases with data from NOAO telescopes, analysis results, and data from external archives in one place
4. Share custom results easily with collaborators, and create and publish catalogs derived from large datasets through a central workspace
5. Experiment with tools being developed for LSST using existing large datasets

The goal of the Data Lab is to provide a common framework and workspace for science collaborations and individuals to use and disseminate data from large surveys, using the best available tools. It is intended to grow into a clearinghouse for knowledge on best practices for handling specific large datasets, and for sharing software developed in the community to work on these datasets in an Open Source manner.

2.2 Project Success Metrics

In broad terms, the project will be a success if individual scientists and survey teams are more productive and publish better science by using it. Usage statistics (e.g., number of users, data retrievals, service access logs, etc) will provide an indication of user uptake and serve as an early indicator of success.

DL-OCD-2200	Access and usage logs will be kept for all services in the Data Lab. Individual components may implement additional logging specific to their functionality.	**N/A**
DL-OCD-2205	Publications acknowledging use of the Data Lab will be tracked over time as an indication of project uptake.	**N/A**

2.3 Project Constraints

A detailed discussion of project planning and resource issues can be found in the Project Execution Plan (PEP) document.

2.4 Flowdown from Science Requirements

Detailed Data Lab operational concepts are given as lists in tabular form in this document, with an identifier DL-OCD-XXXX. If they can be traced to requirements in the SRD, those requirements are given in the third column with identifiers like DL-SRD-XXXXX which were given in the SRD. If an operational concept can not be directly traced to the SRD, the SRD reference is marked as N/A for not applicable; most such operational concepts are directly related to the establishment of the Data Lab computational infrastructure required to achieve the requirements given in the SRD.

2.5 Science Capabilities

In this section, we describe the capabilities derived in the Science Use Case document at a broad conceptual level. In doing so, we identify three classes of users of Data Lab services:

- **Expert users:** Users who are well versed with the catalog database, know exactly the subsample they want, and want to make a query and then proceed directly to custom analysis
- **Novice and intermediate users:** Users who want to explore the catalog database, visualize aspects of it, and then proceed to writing queries that result in custom analyses
- **Collaborations:** A mix of novice, intermediate, and expert users who want to interact with the catalog database in different ways, and save and share output at all stages of the analysis

Our concept is that all of these users will begin in some way with a catalog. The ultimate goal of the initial interaction with the catalog is for the user to design a query to select a subsample, and then perform analysis on it. We will consider two kinds of catalogs:

- Object catalogs containing all data associated with unique positions on the sky
- Source catalogs containing all measurements of flux in peaks above noise on individual frames

While the object catalogs will provide the bulk of the science, the availability of source catalogs will be important for users who need to create custom object catalogs (e.g., moving object catalogs), need to model measurement errors or efficiencies (e.g., through artificial star tests), or wish to perform data quality analysis on a per-frame basis. The SMASH, Bulge, and DES datasets will deliver both object and source catalogs. However, a feature of the Data Lab will be to allow users to create their own object catalogs through a customizable cross-matching service, or through filtering of the large catalog based on some criteria.

The most basic usage of the Data Lab that would support the workflows described in the SUC is for the user to select a catalog, perform a customized query, store the result, and perform custom analysis on it. This mode of operation, which is appropriate for expert users, will be supported by the Data Lab through its database query interface and its user-defined analysis environment. However, we expect that many scientific discoveries will require an initial exploration of the data, particularly for users who are not intimately familiar with the available catalogs. Such data exploration is not best served by forcing the user to make an immediate complex query of the database. For this point of entry, the Data Lab will aid the user with tools for interactive visualization and basic analysis of the catalogs. As much as possible, these interactions with the catalogs will be recorded as database queries, such that they can be reproduced as parts of later quantitative analysis.

Another use of the Data Lab will be help coordinate large catalog science by collaborations of scientists. Individuals in collaborations will interact with the Data Lab in the same ways as single users, but will be able to store their output and procedures to a shared space set aside for the collaboration.

In the following sections, we describe the capabilities that will be included in the Data Lab in more detail. Some of these capabilities may not appear until later stages of development.

DL-OCD-2500	Data Lab will provide access to catalogs using a standard SQL query via command-line tools (for experienced users), Web-based interfaces (for intermediate and novice users) and programmatic interfaces (for application developers).	**DL-SRD-21000** **DL-SRD-21002**
DL-OCD-2502	Data Lab will provide access to catalogs using standard VO interfaces via command-line tools (for experienced users), Web-based interfaces (for intermediate and novice users) and programmatic interfaces (for application developers).	**DL-SRD-21000** **DL-SRD-21002**
DL-OCD-2504	Users will be able to do a positional cross-match of objects in a personal list with catalogs available from the Data Lab or elsewhere. In some cases, local data will be uploaded to a remote service for matching. Lists of non-matches will be available to the user.	**DL-SRD-21700** **DL-SRD-21706**
DL-OCD-2506	Users will be able to sub-select data from large catalogs to create a custom database of results for further analysis.	**DL-SRD-21004**
DL-OCD-2508	Users will be able to sub-select data from large catalogs and download results directly to their desktop, or save to virtual storage for further analysis. Results may be retrieved in a variety of formats suitable for analysis with user-supplied tools.	**DL-SRD-21004** **DL-SRD-21154**
DL-OCD-2510	It will be possible to determine statistical properties of columns in a database beyond the basic functions supplied in an SQL query. Further modeling of the data can be achieved with user-supplied analysis code executed either on the desktop or as a Data Lab processing service.	**DL-SRD-21850**

2.5.1 Catalog Query and Analysis Tools

See section 4 below.

2.5.2 Image Query and Analysis Tools

DL-OCD-2520	Data Lab will provide access to image datasets using standard VO interfaces via command-line tools (for experienced users), Web-based interfaces (for intermediate and novice users) and programmatic interfaces (for application developers).	**DL-SRD-21600** **DL-SRD-21250**

DL-OCD-2521	It will be possible to retrieve DES image data (raw and reduced) from the NOAO Science Archive covering a given position.	**DL-SRD-21602** **DL-SRD-21603**
DL-OCD-2522	It will be possible to retrieve image data from external (i.e. non-NOAO) data services for use in a science workflow. Data Lab services (e.g., cutout) may be used to process these data if the host service cannot fully support the request.	**DL-SRD-21250**
DL-OCD-2524	It will be possible to create image cutouts from data accessible by the Data Lab. This includes automated and efficient processing of position and image lists used to generate an arbitrary number of cutouts.	**DL-SRD-21600** **DL-SRD-21604**
DL-OCD-2526	It will be possible to create image mosaics from data accessible by the Data Lab.	**DL-SRD-21603**
DL-OCD-2528	It will be possible to register images obtained with different filters, exposure times, rotation angles, etc.	**DL-SRD-22100**
DL-OCD-2530	It will be possible to do photometry of images. The Data Lab may provide some core capability based on a known tool (e.g., SExtractor), users can optionally run their preferred code (e.g., DAOPHOT, Dophot, etc).	**DL-SRD-21800**

2.5.3 Time Series Tools

DL-OCD-2540	It will be possible to generate light curves from catalogs containing multi-epoch data. In some cases specialized SQL queries of the catalog will be required, Data Lab will provide examples and documentation to describe these queries.	**DL-SRD-21100**
DL-OCD-2541	Data Lab will support a period-finding service to estimate the period of an object with variable flux.	**DL-SRD-21750**
DL-OCD-2542	Data Lab will support statistical analysis tools to determine if the flux of an object varies in time for a given statistical significance. The evidence for the statistical nature of the variability (periodic, aperiodic, random or transient) can also be determined.	**DL-SRD-21850** **DL-SRD-21855**
DL-OCD-2543	Users will be able to apply their own classifier algorithms to light curves created in the Data Lab.	**DL-SRD-22300**
DL-OCD-2544		
DL-OCD-2545		

2.5.4 Visualization Tools

DL-OCD-2550	Users will be able to display image data in the Data Lab using common desktop tools (e.g., Aladin or DS9) or web applications.	**DL-SRD-21450**
DL-OCD-2551	Users will be able to create arbitrary (i.e. a user-selectable set of columns, expressions or values) 2-D and 3-D plots of tabular data.	**DL-SRD-21400**
DL-OCD-2552	Users will be able to generate phase-folded light curves for a given period from tabular or time-series data. A workflow involving multiple Data Lab components may be used in this process (e.g., data query, compute services and plotting utilities) to build a scripted task.	**DL-SRD-21500**
DL-OCD-2553	It will be possible to create animations of variable objects. Examples include animated GIFs of image cutouts, or full time-ordered data cubes for use by desktop tools.	**DL-SRD-21550**
DL-OCD-2554	Users will be able to easily generate Color-Magnitude or Hess Diagrams (with optional contour overlays) from catalog data to enable graphical analysis of possibly millions of data.	**DL-SRD-21400**
DL-OCD-2555	Users will be able to query for image data using graphical or tabular visualization tools.	**DL-SRD-21550** **DL-SRD-21250**
DL-OCD-2556	Users will be able to upload local files to the Data Lab for visualization and analysis.	**DL-SRD-21153**

2.5.5 User-Defined Analysis

DL-OCD-2570	Users will be able to run their own analysis codes in the Data Lab.	**DL-SRD-22300** **DL-SRD-22000**
DL-OCD-2571	Legacy software will be able to take advantage of programmatic interfaces that can be used to access data and interoperate with Data Lab components.	**DL-SRD-21800** **DL-SRD-22500**
DL-OCD-2572	Legacy software can easily be bundled and deployed in the Data Lab framework as a new compute service.	**DL-SRD-21800** **DL-SRD-22500** **DL-SRD-21900**
DL-OCD-2573	Users will be able to publish reference data for personal use within the Data Lab. For example, a private database of reddening values of a specific part of the sky can be created for use in a workflow. These data can be accessed using standard Data Lab tools.	**DL-SRD-21156** **DL-SRD-21157**

3 System Infrastructure

This section gives a high-level overview of the Data Lab system infrastructure: which components are provided as core capabilities, which are accessible to users for modification, what hardware is available, etc.

3.1 Large Catalog Support

The Dark Energy Survey (DES) catalog will be made available as a first-release midway through the DES survey (currently expected January 2017), and then as a final product once the DES is complete. Additionally, both the yearly releases of the reduced image frames and the raw image/calibration data will be available from the NOAO Science Archive (NSA), giving the Data Lab the ability to query the catalog and then drill-down to the stacked (or reduced, or even raw) pixels as needed. The use of virtual storage in the Data Lab will permit users to save query results to disk formats that may be shared with collaborators or used in analysis workflows. Multiple interfaces to the data will be used to support the variety of clients expected in the most optimal manner. Infrastructure required to support the DES catalog will be used to serve other extremely large catalogs in the future leading up to LSST operations.

DL-OCD-3100	Owing to its size, the DES catalog will require a distributed database system to support parallelized queries in the most efficient manner.	**DL-SRD-2100N**
DL-OCD-3105	Users will be able to save results of DES catalog queries to a personal database table (similar to the SDSS *MyDB* system) for further analysis.	**DL-SRD-21050** **DL-SRD-21055**
DL-OCD-3110	Users will be able to save results of DES catalog queries to a disk file (*allowed formats TBD*) in their virtual storage.	**DL-SRD-21050** **DL-SRD-21055**
DL-OCD-3111	Users will be able to immediately retrieve results of a DES catalog query (*allowed formats TBD*) to their desktop.	**DL-SRD-2100N** **DL-SRD-21050**
DL-OCD-3115	The DES catalog received from the DES Collaboration may require additional indices, computed data columns or table structure to provide optimal response times to support science cases. The final database will be optimized using this added information if necessary.	**DL-SRD-21004**
DL-OCD-3120	Experimental tuning of the database will be conducted on pre-release data prior to operational deployment in order to achieve best *first-byte* performance. If a preview of the DES catalog cannot be obtained then similar catalogs (e.g., from the SMASH pipeline) will be used for development.	**N/A**

DL-OCD-3125	Users will be able to submit DES catalog queries for execution as a synchronous job. Synchronous jobs will be subject to a resource-specific limit on execution times before being aborted.	**DL-SRD-21008**
DL-OCD-3130	Users will be able to submit DES catalog queries for execution as an asynchronous job. Asynchronous jobs will be subject to a resource-specific limit on execution time before being aborted.	**DL-SRD-21006**
DL-OCD-3135	The status of an asynchronous job may be queried by client applications. Optionally, users can elect to be notified (e.g., via email) when a job has completed.	**DL-SRD-21006**
DL-OCD-3140	Users will be notified if any submitted job exceeds its execution time limit.	**N/A**

3.2 Virtual Storage

Virtual storage is used in the Data Lab for a number of purposes:

- as user storage co-located with the Data Lab services that may be invoked,
- for staging results in a workflow before download,
- for uploading input data from the user's desktop,
- as shared storage for all members of a collaboration,
- for publishing data through standard VO protocols for use in workflows,
- for providing alternate views of data for use in applications,

Users will interact with the storage space using web browsers, command-line tools or programmatically, and in some cases may not even be aware it is being used. This space will (appear to) function in much the same way as a regular filesystem, allowing the user to create new directories, allow/restrict access to certain data, and permit standard file operations such as *copy*, *move/rename*, and *delete*. Virtual storage can be mounted remotely as a filesystem on the user's desktop, allowing the user to interact with the space as if it were a local disk (assuming a client *FUSE filesystem*[1] capability).

DL-OCD-3200	Virtual storage will be co-located with Data Lab compute/analytical services for efficient processing. This does not require storage and compute services to be deployed on the same physical machine, only that they are co-located within the Data Lab system.	**DL-SRD-21150**
DL-OCD-3205	Data in virtual storage will be shareable with collaborators via a *user-controlled* mechanism to grant/deny access permission on files being stored to other users.	**DL-SRD-21156** **DL-SRD-21157**

[1] `http://fuse.sourceforge.net/`

DL-OCD-3210	Where possible, virtual storage will permit client applications to request an alternate format of the data stored e.g., a different image format, a compressed file, etc. Not all conversions will be supported; the list of supported views is *TBD*.	**N/A**
DL-OCD-3215	Virtual storage will have a *capability* to allow processing of data moved to the space to be performed on ingest. This may involve backend processing of data to compute alternate formats, the ingestion of data into a publication service, or arbitrary processing of the data such as execution of some task that uses the new file as input.	**N/A**
DL-OCD-3220	Users will be able to control whether and which capability is enabled on a particular storage container (i.e. directory).	**N/A**
DL-OCD-3225	The automated processing mechanism will be general enough to allow a *user-defined* application to be executed on the data (e.g., execute a task stored in the space). This capability requires the user to provide all needed input (e.g., parameter files) for the processing to succeed.	**N/A**
DL-OCD-3230	Users will be able to make the data in their storage containers accessible via standardized data access protocols. We assume this to be a non-public data access interface and those clients will have privileged access to this data using the same credentials used for access to raw file using virtual storage interfaces.	**N/A**
DL-OCD-3235	Users and client applications will be able to easily access and manipulate the contents of a user's Data Lab virtual storage space from their desktop.	**DL-SRD-2115N**
DL-OCD-3240	Users and client applications will be able to access multiple instances of virtual storage.	**N/A**
DL-OCD-3245	Users will be able to download and install the virtual storage service software on their desktop in order to make the same functionality available to local data.	**DL-SRD-2115N**
DL-OCD-3250	It will be possible to synchronize the contents of two or more virtual storage instances.	**N/A**
DL-OCD-3255	An individual user's storage will be categorized as *Personal Space* and will be subject to resource quotas. Data in these spaces may not be backed up in case of hardware failure due to resource limitations.	**N/A**
DL-OCD-3260	Users can apply to have a *Collaborative Storage* space allocated on behalf of a group. This space provides redundant backup and is intended for long-term use by a specific scientific collaboration (e.g., a survey team). The space is owned by a *group* user account that must grant access permission to	**DL-SRD-21156**

	members of the group.	
DL-OCD-3265	A *Transient Storage* space will be available to all users for staging of results from processing or query services. Data in this space will be purged automatically following some (*TBD*) expiration period. Data in this space does not count against an individual user's personal quota.	**N/A**

3.3 Data Publication

The Data Lab will allow individual PIs and Survey Teams to publish their higher-level data products for public access in a number of ways: as a "*simple*" data access service featuring primarily a positional search that is available for catalogs, image or spectra, or through an advanced table access protocol that permits full-access SQL queries of complex catalog datasets. Additionally, users can construct their published datasets so they link easily to the original raw or pipeline-processed data available from the NOAO Science Archive. The intent of this mechanism is to provide a platform for the long-term curation and distribution of NOAO data that don't otherwise fit within the capabilities offered by the NSA. In some cases, these publishing services can be used internally as part of a science workflow.

DL-OCD-3300	Users are responsible for supplying data and metadata in a format that meets the guidelines and requirements of the Data Lab publication scheme.	**N/A**
DL-OCD-3305	Users will be able to publish their image, catalog and spectral data for public access using standard protocols.	**DL-SRD-21156**
DL-OCD-3310	Users will be able to publish complex catalogs containing multiple tables in a way that permits users to query using the entire dataset.	**DL-SRD-21004**
DL-OCD-3315	Simple data collections can be published in a way that permits a parameterized query of spatial, temporal and bandpass attributes of the data.	**N/A**
DL-OCD-3320	Higher-level data products can be connected to the raw or pipeline-processed data that created them.	**N/A**
DL-OCD-3325	Data services can be deployed for private use within the Data Lab with no loss of functionality.	**N/A**
DL-OCD-3330	A private data service can be easily published as a public service.	**N/A**

3.4 Processing Services

The Data Lab will support a limited set of *core processing services* that can be applied to all data available to the astronomer. These services provide general capabilities needed in many science workflows, e.g., image cutouts when processing pixel data, quick photometry of images, statistics of tabular data, etc. Services can, in some cases, be extensible by the user to build a custom capability, but more often it will be preferable to construct a new service built from the core code.

DL-OCD-3400	Processing services will be able to interoperate with the Data Lab virtual storage system, e.g., to use files there as input or for use when storing results.	**N/A**
DL-OCD-3405	The business logic of the processing service can be deployed a number of ways in order for it to be accessible to web, desktop and programmatic applications. The methods used to do this will be shared amongst all services to avoid redundant development for each service.	**N/A**
DL-OCD-3410	It will be possible to invoke a service in a synchronous manner.	**DL-SRD-21008** **DL-SRD-21601** **DL-SRD-21702** **DL-SRD-21752**
DL-OCD-3415	It will be possible to submit a call to a service for asynchronous execution. Clients doing this must be aware of methods for checking on status and returning results.	**DL-SRD-21006** **DL-SRD-21600** **DL-SRD-21700** **DL-SRD-21750**
DL-OCD-3420	It will be possible to deploy multiple instances of a processing service on Data Lab hardware, e.g., to parallelize processing of lists.	**N/A**
DL-OCD-3425	Users will be able to create custom processing services for their own use.	**N/A**

3.5 Machine and Process Virtualization

The DES catalog and other static resources (e.g., virtual storage) will require dedicated hardware to ensure optimal performance, however other aspects of the Data Lab such as support for concurrent users, processing services, or data accesses will be highly variable in terms of resources required. For example, even though the number of supported users might be capped in some way, it is unlikely that they will all be using the system simultaneously and so it is possible to over-subscribe the physical hardware by not provisioning for the worst-case scenario.

DL-OCD-3500	The Data Lab will use virtual machines for components having numerous dependencies on installed system software to operate (e.g., data access services needing databases, web components and physical storage).	**N/A**

DL-OCD-3505	Virtual machines deployed in the Data Lab can be redeployed to new hardware as needed by operations personnel to balance system loads.	**N/A**
DL-OCD-3510	The Data Lab will maintain a set of base operating system images to be used as standard configurations for virtual machines intended for a specific purpose.	**N/A**
DL-OCD-3525	The Data Lab will use lightweight *Linux Containers*[2] for those components that can be isolated easily from underlying system resources (e.g., a single instance of a processing service), or which have high-availability requirements (e.g., a user shell). *Linux Containers are explained in more detail in SAD Sec 1.4.5 .*	**N/A**
DL-OCD-3530	It will be possible to execute multiple instances of a container as a means of increasing capacity as needed.	**N/A**
DL-OCD-3535	Users will be able to download and execute containers on their own hardware as a means of running data lab services locally. Similarly, users will be able to create containers locally and upload them for execution in the Data Lab.	**N/A**
DL-OCD-3540	Linux containers providing shell access to the Data Lab will have access to a user's persistent storage.	**N/A**

3.6 User/Group Authentication

The Data Lab will allow users to create or utilize resources associated with their identity, e.g., virtual storage of data or a personal database. These resources have to be rationed amongst all users based on the hardware available in the system. Users will also expect that any data they put into the Data Lab will ***only*** be available to other users with their explicit permission. Similarly, collaborations will wish to define members of a group and might also choose to restrict or grant permissions to certain members of that group. Data Lab will use a secure authentication mechanism to allocate resources to user identities and to enforce access restrictions.

DL-OCD-3600	Public Data Lab services shall be accessible without requiring authorization.	**N/A**
DL-OCD-3605	A valid NOAO identity can be used to create a Data Lab identity.	**N/A**
DL-OCD-3610	A Data Lab identity can be created using a web portal.	**N/A**
DL-OCD-3615	A Data Lab identity can be created without requiring an NOAO identity.	**N/A**
DL-OCD-3620	The login portal will allow users to see the details of their	**N/A**

[2] `http://en.wikipedia.org/wiki/LXC`

	profile and change information (e.g., password or contact email) as needed.	
DL-OCD-3625	NOAO and Data Lab identities will be connected when both exist; the appropriate credential will be presented to services that require authentication.	**N/A**
DL-OCD-3630	Users will be able to create '*groups*' of other users and use these to set read/write permission on their resources	**DL-SRD-21156**
DL-OCD-3635	Users will be able to set permissions independently on each resource (e.g., storage, database, data service) they control.	**DL-SRD-21156**
DL-OCD-3640	Users will be able to set access permissions using the names of individuals or groups.	**DL-SRD-21156**
DL-OCD-3645	User groups can contain other user's groups as a way to authorize access without requiring individual user names.	**N/A**
DL-OCD-3650	Users who create a group can transfer their *admin* role of the group to another member of the group.	**DL-SRD-21156**
DL-OCD-3640	Data Lab Operations staff have the ability modify the resource quotas assigned to each user.	**N/A**
DL-OCD-3645	Users with an existing NOAO identity do not, by default, have a Data Lab identity.	**N/A**

3.7 Programmatic Interfaces

While the Data Lab will provide a great deal of infrastructure, it will not (specifically) provide the analysis tools required to meet all the stated science requirements. Instead, we assume the bulk of the science code will come from the users themselves. To better support this model, we will package standard tools (e.g., SExtractor) where possible for use in high-level scripting environments, and provide programmatic interfaces (APIs) to data and compute services that can be used when building lower-level applications. These interfaces, combined with interactive shell environments that support a variety of user-specified programming environments (e.g., C/C++, Fortran, Python, R and possibly IDL with a user-supplied license), will provide a rich platform for scientists to develop applications and processing pipelines within the Data Lab or on their desktop.

DL-OCD-3700	*Standard interface* libraries (e.g., http, FTP, database access, graphics) will be available in Data Lab interactive environments as part of the development environment. Users are free to import additional packages as needed.	**N/A**
DL-OCD-3705	*VO protocol interfaces* will be available in Data Lab interactive environment as part of the development environment for use with both internal Data Lab services and external providers.	**N/A**

DL-OCD-3710	*Specific interfaces* to core Data Lab components will be available in Data Lab interactive environment as part of the development environment.	**N/A**
DL-OCD-3715	Users are free to import additional packages as needed into their Data Lab interactive environment.	**N/A**
DL-OCD-3720	All interface libraries available for use in the Data Lab will be available for download and installation to the user's desktop system.	**N/A**
DL-OCD-3625	Data Lab will not mandate that a specific programming language be used for application development.	**N/A**
DL-OCD-3630	Data Lab will not guarantee that a specific language is supported by all interfaces.	**N/A**
DL-OCD-3635	Data Lab may provide language-specific interfaces for platforms commonly used by astronomers.	**N/A**
DL-OCD-3640	Programmer's documentation will be available for all interfaces.	**N/A**

3.8 External Interfaces

Not all data or processing required for the science can be hosted in the Data Lab; this section describes how components will interact with external tools and services, e.g., remote data archives or legacy software on the desktop. Client-side interfaces to these external resources will need to be built into some Data Lab components, or as tools used in the Data Lab infrastructure.

3.8.1 VO Services

All of the major data centers provide standard VO data access interfaces to their major holdings, and centers such as CDS[3] provide name resolver and cross-match catalog services in addition to their large collection of catalogs from published papers. Data Lab will be able to make use of these services in scientific workflows as well as interoperate with similar services (e.g., virtual storage) hosted elsewhere.

DL-OCD-3800	*Data Lab applications*[4] will be able to query for data at remote centers using standard VO protocols. This applies to web, desktop and programmatic tools developed in the Data Lab; it is optional for user-developed tools.	**N/A**

[3] `http://cdsweb.u-strasbg.fr`

[4] E.g. processing services or command-line tools that might be run from a user shell, or any Data Lab component needing to make VO client-side requests of a service.

DL-OCD-3805	*Data Lab applications* will be able to invoke compute services such as name resolvers, catalog cross-matches, image cutouts or other specialized tools hosted at remote data centers. This applies to web, desktop and programmatic tools developed in the Data Lab; it is optional for user-developed tools.	**N/A**

3.8.2 Desktop Tools

Existing purpose-built desktop tools provide tremendous functionality that interacts equally well with local (i.e. on the user's machine) and remote data and often interfaces with legacy analysis systems as well. Data Lab will be able to serve data to these tools using a variety of interfaces and allow the user to choose their preferred tools for interacting with the data.

DL-OCD-3825	Web-based applications (e.g., catalog query forms) will be able to broadcast messages to desktop tools for the purpose of loading data in the tool.	**N/A**
DL-OCD-3830	Data Lab will serve data through the standard interfaces supported by commonly used desktop applications (e.g., Topcat, Aladin, DS9).	**N/A**
DL-OCD-3835	Programmatic interfaces developed in the Data Lab will be able to interface to desktop tools to load and manipulate the data.	**N/A**

4 Science Operations Overview

This section provides a more detailed discussion of the scientific use of the Data Lab, e.g., the practicalities of querying large catalogs or constructing workflows using local and remote data sources. It is written from the perspective of the science user.

4.1 Large Catalog Queries

Even though the DES Catalog (and others) will be optimized for the most common use cases, complex queries can still be expected to occasionally require a long time to complete. Users, then, cannot assume that all queries will be interactive, but instead may prefer to submit the job and have the results staged to their virtual storage or personal database (*MyDB*). Cross-match jobs in particular may take a long time to execute if the matched catalog is not locally stored in the Data Lab or if the entire DES database is to be searched.

DL-OCD-4100	Users will be able to submit queries to a batch-processing queue for execution.	**DL-SRD-21006**
DL-OCD-4105	It will be possible for users or client applications to check on the status of queries to determine when the job is complete.	**DL-SRD-21006**
DL-OCD-4110	Users will have the option of being notified by the system when a batch job is complete.	**N/A**
DL-OCD-4115	Users will have the option of saving query results to a personal database when a job is complete. The user must specify this option (e.g., with a *SELECT INTO* clause in the SQL) when submitting the query.	**DL-SRD-21050** **DL-SRD-21055**
DL-OCD-4120	Users will have the option to save query results in a file in their virtual storage area.	**DL-SRD-21050** **DL-SRD-21055**
DL-OCD-4125	Users will be able to upload a local disk file, or file in virtual storage, to be used in a cross-match of the DES catalog.	**DL-SRD-21153**
DL-OCD-4130	Users will be able to submit queries to a synchronous service for immediate processing. The execution time will be capped to some TBD fixed length of time to provide predictable response times for client applications. The return of partial results is not guaranteed in cases of execution timeout.	**DL-SRD-21008**
DL-OCD-4135	Users will be able to query the DES catalog using desktop tools that use standard VO protocols for data access.	**DL-SRD-21002**
DL-OCD-4140	Users will be able to query the DES catalog using custom tools that operate using Data Lab interfaces.	**DL-SRD-21000**

4.2 Visualization

Visualization is an important part of understanding both the dataset as a whole, and the analysis that was done on a subset. Interacting with a plot, spectrum or image is also sometimes required to provide input to an application or to get more detailed information (e.g., a value under the cursor of a plot or image). The Data Lab will provide core tools for visualizing data of all types; these *may* be developed as toolkits that could then be extended for science-specific plotting applications. When appropriate, multiple views of the data will be available, e.g., a 2-D plot may show the tabular data represented along with an image of the search field in a catalog query. Interoperability with purpose-built desktop tools having more graphics capability will also be supported as appropriate.

4.2.1 2-D Plotting

The core plotting tools in the Data Lab will allow the user to create generalized 2-D plots constructed from individual columns or arbitrary expressions of columnar values, e.g., an expression creating a color value from single-filter magnitude columns, or the deviation of some value from the mean value of the column. The Data Lab will need graphics capabilities in a number

of places depending on the application being used: in desktop tools that may have built-in graphics but no specific knowledge of the Data Lab services, in browser-based applications used to query and analyze data, and in programmatic interfaces use to develop science-specific tools. Color-Magnitude Diagrams or Hess diagrams are instances of a specialized plot types and examples of plots that could be created by the plotting toolkit, however the details of how these plots are generated will differ depending on the use-case.

DL-OCD-4200	Developers should be able to use their chosen graphics package when developing applications using the Data Lab programmatic interfaces.	**N/A**
DL-OCD-4205	A variety of plot types should be supported. These include (but are not limited to): • X-Y / Scatter plots • Histograms • Contour plots • Light curves / Spectra	**DL-SRD-21400** **DL-SRD-21500**
DL-OCD-4210	Plots should be able to make use of error information when available to plot error bars.	**DL-SRD-21400** **DL-SRD-21500**
DL-OCD-4215	Users should be able to control plotting window limit, axis labels and annotation text.	**N/A**
DL-OCD-4220	Vector plots should support a variety of line and fill-area styles with user-selectable colors and/or colormaps.	**N/A**
DL-OCD-4225	Overplotting of graphics should be supported (e.g., contours over density shades).	**DL-SRD-21400**
DL-OCD-4230	Users should be able to plot onto the celestial sphere, ideally with a choice of projection options.	**N/A**
DL-OCD-4235	3-D plotting should be supported, ideally with interactive manipulation of the viewing parameters.	**DL-SRD-21450**
DL-OCD-4240	It should be possible to save any graphic to a disk file. The allowed formats are *TBD*.	**N/A**
DL-OCD-4245	All plotting tools should work efficiently with large datasets. This may require server-side computation of plots (for web-based tools) or other optimizations such as minimizing the display of redundant data within the plot resolution.	**N/A**
DL-OCD-4250	It should be possible to use existing desktop visualization tools in some instances.	**DL-SRD-21400** **DL-SRD-21500**

4.2.2 Image Display

Image display is analogous to the problem of catalog visualization in the sense that one might not always want all of the available data, e.g., a small image display of a cutout (say 4KB in size) from the larger DECam FITS image (~1GB) or a similarly smaller graphic preview of the entire field. As with the plotting service discussed above, it makes sense in this use case to do the computation (i.e. the cutout or the preview generation) in the Data Lab and transfer only the result. At other times it will be necessary for the entire image to be available, e.g., for use in an image display tool such as DS9[5] or Aladin[6] where interactive analysis requires a full image header and actual pixel values.

DL-OCD-4270	It should be possible to display images located in virtual storage using any desktop image display that accepts a URL or local disk file. In some cases, specialized client software will be necessary to load the display.	**N/A**
DL-OCD-4275	It should be possible to display images to a desktop application from a web page listing the results of an SIA query.	**N/A**
DL-OCD-4280	It should be possible to display multiple image cutouts in a web browser.	**N/A**
DL-OCD-4285	It should be possible to display multiple image cutouts of the same object in an animated display (e.g., an animated GIF)	**DL-SRD-21550**

5 Technical Operations Overview

This section provides an overview discussion of the technical operations of the Data Lab and presents concepts describing daily operations required of the developers and operators. This may involve development of utility software not otherwise needed by science users.

5.1 Large Catalog Support

5.1.1 DES Catalog

The DES catalog is scheduled to have only two formal releases, however improved database optimization schemes, hardware upgrades and updates to the underlying software mean that the catalog service itself is likely to be rebuilt more frequently.

DL-OCD-5100	Only Data Lab administrators will have write access to the DES catalog databases, all other users will have read-only access.	**N/A**
DL-OCD-5105	The creation of a Data Lab identity will not be required to	**N/A**

[5] http://ds9.si.edu/

[6] http://aladin.u-strasbg.fr/

	access DES data, however registered Data Lab users may have additional resources available to them (e.g., a personal database to save results).	
DL-OCD-5110	Data Lab will develop utility tools to help automate the deployment of the DES catalog service.	**N/A**
DL-OCD-5105	Procedures used to deploy the DES catalog will be carefully documented and refined with each update of the service. This includes documenting the hardware configuration and tracking of installed software dependencies.	**N/A**
DL-OCD-5120	Procedures for recovering the loss of one or more machines in the system will be developed and tested prior to public release of the DES catalog. This is intended to minimize downtime due to hardware failure without having to rely on redundant systems.	**N/A**
DL-OCD-5125	Data Lab administrators will have tools to monitor the health and activity of the DES catalog services, including the ability to terminate or initialize query processing.	**N/A**

5.1.2 Other Large Catalogs

In addition to the DES catalog, other *Large Catalogs* may someday be available within the Data Lab that also require a distributed database. These may be published data products from large NOAO Survey Programs, or may be high-value reference catalogs (e.g., SDSS or potentially a DESI object catalog) imported to optimize common functionality such as a local catalog cross-match. In principle the steps of partitioning the data across multiple machines, setting up the databases and deploying the front-end services will be similar to what is required for the DES catalog. However, new data sources will have their own schema structure, and if not derived from DECam data, may have a different mapping to the standard data model which will require additional customization of the service.

DL-OCD-5150	Tools developed for deploying the DES catalog will be designed to be easily extensible to working with other large catalogs in the future.	**N/A**
DL-OCD-5155	The effort required to support additional large catalogs will have to be evaluated on a case-by-case basis and scheduled according to other development priorities.	**N/A**
DL-OCD-5160	If not derived from DECam data, other Large Catalogs may require additional customization of the service front-end to be supported.	**N/A**
DL-OCD-5165	Deployment of additional large catalogs will require significant lead-time notice in order to provision hardware resources and modify service and tools prior to release.	**N/A**

5.2 Virtual Storage

A unique feature of the Virtual Storage service we plan to develop is that it will be exportable to run on the user's hardware as well as on Data Lab hardware. On a user's machine, the virtual storage service could be pointed to a local hard disk to act on the user's data directly, allowing the user to take advantage of the service features (e.g., making it searchable via standard protocols, format conversion, etc) with minimal overhead and without requiring that all user data be uploaded into the Data Lab to be made useful. Within the Data Lab, virtual storage would sit atop a large filesystem partition that defines the space available. Individual users have their data stored in subdirectories of the root container and the storage containers they create reflect the directory structure of the space itself, i.e. analysis code can use a '*path*' to the data that is easily derived. Because access to the underlying files is still controlled by the Virtual Storage interface, the service will be able to restrict access to the data to authorized users.

DL-OCD-5200	Virtual storage will use a normal filesystem scheme to minimize overhead of running the service, and to simplify deployment of that service to user hardware	**N/A**
DL-OCD-5205	Users will be able to run a Virtual Storage service on local hardware to manage their files and still take advantage of the service functionality.	**N/A**
DL-OCD-5210	Virtual storage service will be installable by a user with minimal overhead.	**N/A**
DL-OCD-5215	Operations personnel and system administrators will manage the virtual storage disk array according to industry best practices.	**N/A**
DL-OCD-5220	Expanding the storage capacity can be accomplished by adding new hardware and mounted partitions and exposing these to the Virtual Storage service layer.	**N/A**
DL-OCD-5225	Storage quotas imposed on each user can be managed using standard Linux tools.	**N/A**
DL-OCD-5230	Direct access to the disk store will not be permitted by user applications other than through a FUSE-mounted file filesystem.	**N/A**
DL-OCD-5240	This space will serve not just as long-term storage, but also as working storage for use during analysis workflows. The initial implementation of the virtual storage system, then, must work directly with native files of the machine in order to permit modification of files in virtual storage.	**N/A**

5.3 Data Publication

The term *Data Publication* in this section has two meanings: *registering* the metadata for a collection in a central Registry so that the collection may be discoverable via keyword, data type, spatial, temporal or bandpass queries, and *publishing* the data so that it may be accessed using standardized data access protocols. In both meanings, a standard *ObsCore*[7] data model (see summary in Appendix II) is used to describe the data (position on the sky, type of data, responsible PI, etc), the motivation for this model is that it can easily be mapped to concepts common to all datasets regardless of origin. The required elements of ObsCore also serve to define what information must be obtained for the data collection to be discoverable by other users, and for the data to be made accessible.

DL-OCD-5300	Registering a dataset is only required for those collections we wish to make discoverable by outside users, e.g., the DES catalog and the final data products of the SMASH and Bulge Surveys.	**N/A**
DL-OCD-5305	Within the Data Lab it will be possible to create and use data access services without formally registering them with the VO provided that client applications somehow know the service URL	**N/A**

5.3.1 Collection Metadata and Registry Records

In order to register a data collection fully, we must be able to describe all the attributes required in the Registry record. This information is generally available from the originator of the data (e.g., the Survey Team), it does need to be detailed in order to still be fully functional, however as much information will be registered as is possible. Once the information is obtained the registration can be done manually by filling out a web form at the VO Registry being used[8]. Alternatively, data providers may operate a *publishing registry* that maintains the registration records locally; these are harvested regularly using standard protocols[9] to populate a fully-searchable VO Registry located elsewhere.

DL-OCD-5325	Public Data Lab data services (e.g., SMASH, Bulge and DES initially) will be manually registered with an existing Registry service during the development phase of the Data Lab. A local publishing registry will be created once a need to manage more than just a few resources makes this a burden.	**N/A**

[7] Louys, M., et al, "Observation Data Model Core Components and its Implementation in the Table Access Protocol,
Version 1.0", `http://ivoa.net/documents/ObsCore/` , IVOA Recommendation 28 October 2011

[8] For the Data Lab we will be using the VAO Registry at `http://vao.stsci.edu/publishing`

[9] See `http://www.openarchives.org/OAI/openarchivesprotocols.html`

DL-OCD-5330	Data publication tools will be easily extensible to create needed XML registration records when a local publishing registry is eventually required.	**N/A**
DL-OCD-5335	Users wishing to publish data through Data Lab services are required to provide all the necessary collection metadata needed to properly register the service.	**N/A**
DL-OCD-5340	The Data Lab will establish a *naming authority* within the Registry in order to register the resources it provides.	**N/A**
DL-OCD-5245	Data Lab will develop a local *resource resolution* service that can be used to resolve identifiers (e.g., a virtual storage file identifier) to a specific access URL. This will allow identifiers to be used in the Data Lab without requiring that all resources be publically registered (e.g., for temporary, or private, data services).	**N/A**
DL-OCD-5246	The resource resolution service will call a VO Registry to resolve a resource that is not found locally, allowing client applications to use a single service for identifier resolution.	**N/A**

5.3.2 Data Ingestion

The steps required for data ingestion (i.e. putting catalogs into databases, or creating searchable databases of FITS files) depend greatly on the original format of the data and its type (catalog, image or spectrum). Here we describe the process of ingesting higher-level data products to be published as public data services, the process for *Large Catalogs* such as DES is considerably more involved and is described in Sec 4.1. Similarly, data services created automatically, e.g., as a virtual storage *capability*, are described in Sec 4.2.

The development phase of the Data Lab will concentrate on DECam-based surveys, meaning the image metadata will be fairly uniform and simplify the tools required to scan each image to collect the needed metadata. Catalogs created by SMASH/Bulge will be unique and we will leave it to each survey team to define the content. Publication of spectral data is not required during development but will be supported in the full implementation. Data Lab Operations personnel will ingest the data and be responsible for final publication of the service. Because this is a manual process we will not be able to support frequent updates to data such as those that may be produced during an ongoing survey program. Once a dataset is published, there should be no need for ongoing support other than that caused by major changes to the underlying service code (e.g., implementation of a new version of the protocol).

General Concepts

DL-OCD-5350	The Data Lab will provide guidelines on acceptable data formats and metadata requirements; users are responsible for converting their data into one of these data formats.	**N/A**
DL-OCD-5351	Users are responsible for providing the metadata mapping needed to publish the data.	**N/A**

DL-OCD-5352	Data publication will be limited to DECam-based image collections during Data Lab development.	**N/A**
DL-OCD-5353	All datasets to be published must first be delivered to the operator in some way, preferably as a network transfer to/from some FTP staging area. Details will be negotiated with Data Lab Operations personnel.	**N/A**
DL-OCD-5354	Data Lab personnel will load the database, move data files to their final storage location and deploy the data service. Users will not have direct access to Data Lab services used to publish public data.	**N/A**
DL-OCD-5355	New data releases are expected to be infrequent (e.g., yearly) and simply incremental updates to the initial dataset.	**N/A**

Catalog Data Concepts

DL-OCD-5360	Data provided by users must be in a format that can be easily loaded into a database. Formats such as CSV or SQL dumps can be read directly, other formats may require helper utilities such as *STILTS*[10].	**N/A**
DL-OCD-5361	An ingestion tool will read a user-provided configuration file that defines the required service metadata (e.g., primary celestial position columns) in order to create the service-specific configuration file.	**N/A**

Image Data Concepts

DL-OCD-5370	Data provided by users must be in a format that can be served directly to users, e.g., files must pass basic FITS validation and not require special software to interpret correctly.	**N/A**
DL-OCD-5371	Data Lab will provide guidelines on required keyword metadata for each image, including the conventions to be used for (e.g., WCS and FITS) inheritance.	**N/A**
DL-OCD-5372	An ingestion tool will read a user-provided configuration file that defines the required service metadata and scans the image files to construct 1) a searchable database for the service and 2) the service-specific configuration file.	**N/A**

Spectral Data Concepts

DL-OCD-5380	Data provided by users must be in a format that can be served directly to users, e.g., files must pass basic FITS validation and not require special software to interpret correctly.	**N/A**

[10] `http://www.star.bris.ac.uk/~mbt/stilts`

DL-OCD-5381	Data Lab will provide guidelines on allowed spectral formats (tabular formats, native FITS, etc) and required metadata. Users are responsible for converting data to one of these supported formats.	**N/A**
DL-OCD-5382	An ingestion tool will read a user-provided configuration file that defines the required service metadata and scans the spectral data files to construct 1) a searchable database for the service and 2) the service-specific configuration file.	**N/A**

Time-Series Data Concepts

DL-OCD-5385	*Support for Time Series data is TBD*	**N/A**

Spectral Line Data Concepts

DL-OCD-5390	*Support for Spectra Line data (e.g., reference spectral atlas) is TBD*	**N/A**

Ephemeral or Internal Data Concepts

DL-OCD-5395	*Support for private data services is TBD*	**N/A**

5.3.3 Metadata Harvesting

The Data Lab will implement a number of metadata harvesting tools, one for each type of data (catalogs, images, spectra, etc) being published. These tools will collect and map metadata directly from the data files to construct a searchable database that can be used by the publishing service. For each tool, the common pattern is that a user-supplied mapping of the ObsCore *concepts* (e.g., sky position) to the specific *instances* (e.g., header keywords or column name) in the data are used as input to the task along with some specification of the files to be harvested (e.g., a directory name, specific list of files, database connection and table name, etc)

DL-OCD-5396	Harvesting tools will support a limited number of file formats to simplify initial development. For images, support will be limited to data coming from the DECam/Mosaic/NEWFIRM instruments or pipelines.	**N/A**
DL-OCD-5397	Data Lab will provide guidelines and methods for users to specify the metadata mapping needed by the harvesting tools.	**N/A**
DL-OCD-5398	Metadata harvesting is one step in the process of ingesting data for publication; it can be run independently of ingestion as a means of verifying the data conform to Data Lab guidelines and that all required information is present.	**N/A**
DL-OCD-5399	Users will be able to download and install the harvesting tools to locally verify compliance with guidelines before	**N/A**

	transferring data for ingestion.	

5.4 Processing Services

A *processing service* is the generic name given to any compute service run in the Data Lab that does some form of data processing or analysis, e.g., a *cutout* service for images or a *statistics* service for tabular data. It is called a service because it may be deployed any number of times within the system, providing backend computation for web applications or as part of a science workflow analyzing results of a data query. In many cases, the Science Capabilities described in Section 2.5 above will form the basis for these services.

DL-OCD-5400	Compute services will understand and utilize Data Lab interfaces where possible, e.g., to access virtual storage or query large catalogs as part of a workflow.	**N/A**
DL-OCD-5405	These services may be implemented using available desktop tools or legacy systems. By containerizing the service we are able to bundle dependencies and run the service on many types of machines.	**N/A**
DL-OCD-5410	These services may act as clients for other services (e.g., as a client for a remote cross-match or period-finding service).	**N/A**
DL-OCD-5415	It will be possible to access select services from outside the Data Lab framework. Some core services will be generally useful in contexts outside the Data Lab as standard web services; these will be hosted as permanent services that do not require authentication.	**N/A**

5.5 Machine and Process Virtualization

A Virtual Machine might be implemented as dedicated hardware having a large persistent storage and a custom configuration for use by the team members in cases where Data Lab is being used extensively by a particular science collaboration (e.g., the SMASH survey), but more generally VMs will be used as a means to provision hardware for services within the Data Lab. Users will access Data Lab primarily through specialized login containers or by calling a processing service. Containers are accessible in sub-second startup times (as opposed to minutes for a VM to boot), giving *all* users constant access to their environment and eliminating both the need for privileged access and conflicts with other users.

DL-OCD-5500	Users of the Data Lab will have shell-level access by means of a specialized login container; this gives them a configurable system in a container that isolates their environment from	**N/A**

	the underlying system.	
DL-OCD-5505	Users will be able to create containers for software they develop themselves. Data Lab will provide documentation and tools to help users create containerized applications.	**N/A**
DL-OCD-5510	Containers can be deployed as a processing service within a workflow, or may be made available to other users from the repository.	**N/A**
DL-OCD-5515	Containers will have access to a user's virtual storage but may also be configured to have a small working disk allocation mounted as a logical partition.	**N/A**

5.6 User Account and Authorization Management

Users will be able to create a Data Lab login themselves through the web portal, however Operators will need to recover lost passwords, bulk-register or delete users, troubleshoot problems, and otherwise oversee resources associated with a user identity.

DL-OCD-5600	Data Lab operators will be able to create a Data Lab identity from an existing NOAO identity and ensure that either identity will authorize a user.	**N/A**
DL-OCD-5605	Data Lab operators will have tools to create (or delete) user accounts, manage group membership and allocate resource quotas assigned to individual users or groups.	**N/A**
DL-OCD-5610	Data Lab operators will have tools to remove and release all resources associated with a specific user account.	**N/A**
DL-OCD-5615	Operators will have tools to backup and restore user resources deemed as *permanent* within the Data Lab. Users will be responsible for creating backups of all other resources if needed.	**N/A**

5.7 Source Code Repository

The Data Lab will maintain several types of repositories to be used by software developers, astronomers and operations staff. All development will be done in an Open Source manner, however not all material developed is appropriate or necessary to share publically.

DL-OCD-5700	Data Lab Developers will use a version control system for software management.	**N/A**

DL-OCD-5705	All software deployed for use in the Data Lab will be maintained in a *public repository* (e.g., *GitHub*[11]) and available under an Open-Source license (details *TBD*). Core Data Lab software must meet documentation standards described below, user-developed software will be expected to follow similar guidelines but is the responsibility of the developer.	**N/A**
DL-OCD-5710	A *private repository* (e.g., *GitLab*[12]) will be maintained for managing utility software, configuration files, user-sensitive data, etc. This repository will also be available for use by developers or astronomers during the initial development stages of a project.	**N/A**
DL-OCD-5715	Users will be able to build applications by pulling from either the public or private repositories.	**N/A**
DL-OCD-5720	An *archive repository* (e.g., a large local disk) will be maintained for storing data needed for operations staff, e.g., virtual machine images, database backups, etc. Users will not have access to this repository.	**N/A**

5.8 System Documentation

Documentation will be written for all components of the Data Lab; in some cases a component will be described in multiple documents targeted at different levels of user expertise or for different purposes (e.g., software development of a core service versus access to it by a user). This documentation breaks down into the several categories described below.

5.8.1 User Documentation

User documentation will most often take the form of a *HowTo* or *Cookbook* document and is written to instruct the reader (i.e. the user) on how to accomplish a particular task in the Data Lab (e.g., a demonstration workflow), or serve as a more general introductory guide for some particular skill required (e.g., an introduction to SQL, a guide to the DES catalog data model, etc.). Manual pages for individual tasks are also considered user documentation and will be required for all released software.

DL-OCD-5800	Both technical and scientific personnel, as appropriate for the subject, will write documentation.	**N/A**
DL-OCD-5805	All end-user applications and services will have appropriate user documentation (e.g., a manual page or cookbook) describing its use.	**N/A**

[11] `http://www.github.com/`
[12] `http://about.gitlab.com/`

DL-OCD-5810	Community-written documents (e.g., a description of a scientific workflow) will be made available to other users through the document repository and/or online.	**N/A**
DL-OCD-5815	Web-based tools will contain context-sensitive tooltips on mouse-over.	**N/A**
DL-OCD-5820	All desktop tools will support a "—help" option containing examples.	**N/A**

5.8.2 Developer Documentation

Developer documentation is written to document the interfaces or structure of a software system or service. Unlike user documentation, the bulk of these documents can be generated automatically from the source code itself using systems such as *Sphinx*[13] (for Python), *Javadoc*[14] (for Java) or *Doxygen*[15] (for C/C++ and others). An exception to this is the documentation required for the parameters and methods of a data or compute service that a user might call (e.g., a cross-match or cutout service) from a scripting environment. Developer documentation also includes a test plan for the particular application or service.

DL-OCD-5830	The programmer responsible for the software will also be responsible for generating the developer documentation.	**N/A**
DL-OCD-5835	Documentation generators such as *Sphinx, Javadoc,* or *Doxygen* will be used when implementing all public interfaces.	**N/A**
DL-OCD-5840	User-facing compute or data services will be documented to describe: • how the service is called • a full description of all required and optional input parameters • a full description of what is returned by the service, including possible error returns	**N/A**
DL-OCD-5845	Specialized documentation will be written, as needed, describing the interaction of Data Lab components with common desktop tools (e.g., Topcat, Aladin, DS9). These will document, for instance, the messaging interface supported by the tool and how it might be used in a particular use case, or the steps required in a tool to interact with the Data Lab (e.g., to query a local data service).	**N/A**

[13] http://sphinx-doc.org

[14] http://en.wikipedia.org/wiki/Javadoc

[15] http://www.stack.nl/~dmitri/doxygen/

5.8.3 Operator Documentation

Operator documentation is intended to describe how to operate software in the Data Lab, how to deploy new instances, troubleshoot problems, manage hardware resources and recover from failures. In cases where software or services are available for user download; the necessary operator documentation will be included with the user documentation for the software/service. Because this documentation is intended for internal project use, posting it to the project TWiki will help ensure it can be updated easily.

DL-OCD-5850	The programmer responsible for the software will work with the Operations staff to help develop appropriate operator documentation.	**N/A**
DL-OCD-5851	Documentation will be written to keep track of how services are deployed across all available Data Lab hardware.	**N/A**
DL-OCD-5852	Documentation will be written describing how to install and configure each service available in the Data Lab.	**N/A**
DL-OCD-5853	Documentation will be written describing how to manage users within the Data Lab, e.g., how to manually add individuals or groups, how to reset lost passwords, remove users, or modify quotas assigned to them.	**N/A**
DL-OCD-5854	The software dependencies of each deployed service will be documented, e.g., required version of a language, required libraries, etc.	**N/A**
DL-OCD-5855	Procedures for monitoring the health and correctness of systems will be documented.	**N/A**

5.8.4 Document Repository

All project documentation will be available from the project website as well as in the main code repository.

DL-OCD-5860	All system, user and developer documentation will be available from a central documentation repository.	**N/A**
DL-OCD-5861	The contents of the document repository will be visible on the project's website for public access.	**N/A**
DL-OCD-5862	The project TWiki will be used as a repository for documents in-development or for internal use only. Access to this area of the TWiki will be restricted to developers.	**N/A**
DL-OCD-5863	Users will be able to contribute documents to the repository as appropriate. For example, a cookbook document for a science workflow should be available from the primary document	**N/A**

	repository without requiring operator intervention.	

5.9 System Operations and Monitoring

Data Lab operators will need methods to monitor system health and loading to ensure all components are available and running properly. These tools will generally not be visible to users, however users will be able to monitor their use of Data Lab resources from their portal login page.

DL-OCD-5900	Data Lab will automatically monitor the health of hardware in the system. These checks will also alert users to unusual system loads or performance problems.	**N/A**
DL-OCD-5905	Periodic checks of data services will be performed automatically to ensure services are available and responding.	**N/A**
DL-OCD-5910	Usage reports will be generated from system logs and will be available from a status web page visible to all users.	**N/A**

6 User Interfaces

This section describes various concepts being considered for the different user interfaces. Users will interact with the Data Lab in a number of ways, ideally the same tool will be available from each of the interfaces without requiring it be re-implemented.

6.1 Web interface Concepts

DL-OCD-6100	Users will be able to browse their data in the Data Lab from a web interface, specifically: • Virtual storage holdings • Private data service interfaces • Personal database tables	**N/A**
DL-OCD-6105	It will be possible to create *connected views* of the data, for example a web page containing an image cutout, a 2-D plot and a tabular listing where selecting an area of the plot highlights corresponding rows in the table and overlays marks on the image cutout for those points.	**N/A**
DL-OCD-6110	A standard web page interface will be available for all published data services.	**N/A**
DL-OCD-6115	A custom web page will be available for DES catalog queries. This page will allow the creation of SQL queries and monitoring/control	**N/A**

	over asynchronous jobs.	

6.2 Command-line Tool Concepts

DL-OCD-6200	Applications may be constructed from Data Lab interfaces as well as 3rd party software. Additionally, tools can be written using the most appropriate language.	**N/A**
DL-OCD-6205	Tools developed within the Data Lab will interface easily to both local and remote data sources. Remote input data may be specified by a URI (e.g., a URL or Virtual Storage identifier).	**N/A**
DL-OCD-6210	Applications will be able to transparently access Data Lab resources that require authentication (e.g., a user's virtual storage).	**N/A**
DL-OCD-6215	It will be possible to run a command-line tool in a linux container. This allows a tool be used on machine that might not have the tool's dependencies installed, or to be run in a cluster environment.	**N/A**
DL-OCD-6220	It is possible to write a high-level application using the "*tasking*" interface; this provides a front end to operate as a desktop tool but also allows the task to be called from a programmatic interface and easily pass parameters and retrieve output.	**N/A**

6.3 Programmatic Interface Concepts

DL-OCD-6300	A *tasking interface* will be developed to make non-interactive desktop applications callable from a programmatic interface. This serves as the primary interface to legacy code with high-value analysis capability or to containerize processing services.	**N/A**
DL-OCD-6305	Select APIs will be developed (or exist) as C/C++ interfaces. Language bindings for these interfaces will be auto-generated (e.g., with SWIG[16]) to support the most common development languages with a single code base.	**N/A**
DL-OCD-6310	Use of *remote objects* in Python code (e.g., packages such as PyRO[17] or RPyC[18]) will be investigated. These systems would permit local execution of a script that acts on a remote data object such as an image/table/spectrum in the Data Lab.	**N/A**

[16] `http://www.swig.org`
[17] `http://pythonhosted.org/Pyro4/`
[18] `http://rpyc.readthedocs.org/en/latest/`

6.4 User Logins

DL-OCD-6400	Users will have shell-level access to the Data Lab via a specialized login container, allowing them to develop tools and workflows in an environment with efficient access to the data.	**N/A**
DL-OCD-6405	Collaborations will be able to create a common group login identity that may be shared amongst members of the collaboration.	**N/A**
DL-OCD-6410	Users will be able to configure their shell environment by installing new software or uploading data. This environment is peculiar to that user and will persist until the user chooses to reinitialize to the default shell. User shells can be saved in order to persist the environment, they can likewise be used as a new base image for other users.	**N/A**

6.5 Help Desk and Community Engagement

DL-OCD-6500	Data Lab will provide support to users using a number of methods: • A contact email address for help resolving problems • User forums or some comparable system • An issue tracking system to report bugs in specific tasks	**N/A**
DL-OCD-6505	Except for areas where sensitive information (e.g., user data, security issues) might be discussed, the Data Lab wiki will be visible to all users from the project web site. Procedures will be established to determine how/when community users may edit the wiki themselves. *In preparation for the Data Lab Conceptual Design Review, the development team has used an internal Data Lab TWiki for the sharing of documents and information.*	**N/A**
DL-OCD-6510	A Data Lab project website will be available as a resource to find information or software, and as an entry portal for registered users. Procedures will be established to determine how/when new registrations are approved.	**N/A**
DL-OCD-6515	Data Lab resources will be available to any astronomer granted time on NOAO facilities. Efforts will be made to encourage use of the Data Lab by survey teams.	**N/A**

Appendix I: Vocabulary / Acronyms Used

AAS	(American Astronomical Society)
ADASS	(Astronomical Data Analysis Software and Systems) conference
ADQL	(Astronomical Data Query Language) An SQL-like language which includes astronomical facilities to query a database.
AGN	(Active Galactic Nucleus)
API	(Application Programming Interface) The documentation of the interface to a software library or tool.
ASCII	(American Standard Code for Information Interchange) A character-encoding scheme based on the English alphabet where 128 specific characters are encoded into 7-bit binary integers.
ASV	(ASCII Space Values)
AURA	(Association of Universities for Research in Astronomy)
CADC	(Canadian Astronomy Data Centre)
CDS	Centre de Données astronomiques de Strasbourg
CMD	(Color Magnitude Diagram)
CSV	(Comma Separated Values)
CTIO	(Cerro Tololo Inter-American Observatory)
DAL	(Data Access Layer) The VO protocols that define how VO applications access data resources.
Datalink	VO protocol for associating complex astronomical data
DECaLS	DECam Legacy Survey
DECam	(Dark Energy Camera) A 520 megapixel digital camera on the Blanco 4-m telescope at CTIO.
DES	(Dark Energy Survey) a survey to prove the origin of the accelerating Universe and help uncover the nature of dark energy by measuring the 14 billion-year history of cosmic expansion with high precision over five years beginning in summer 2013.
DESI	(Dark Energy Spectroscopic Instrument) An instrument to measure the effect of dark energy on the expansion of the universe by obtaining optical spectra for tens of millions of galaxies and quasars (beginning 2018).
DESDM	(Dark Energy Survey Data Management) Project that developed and operates the DESDM system at NCSA.
DL	(Data Lab)
DAOPHOT	Package for crowded field stellar photometry.
Docker	An open platform for developers and system administrators to build, ship, and run distributed applications.
DoPHOT	CCD PSF fitting photometry program.
DS9	SAOimage DS9, an astronomical imaging and data visualization application.
DSS	(Digitized Sky Survey)
ESO	(European Southern Observatory)
FITS	(Flexible Image Transport System) An open standard defining a digital file format for storage, transmission, and processing of astronomical (and other scientific) data.
FTP	(File Transfer Protocol) A standard network protocol used to transfer computer files from one host to another host over a TCP-based network.

FUSE	(FileSystem in User Space) An operating system mechanism that lets non-priviledged users to create their own file systems.
GAVO	(German Astrophysical Virtual Observatory)
GMS	(Group Management Services)
GPFS	(General Parallel File System) A high-performance clustered file system developed by IBM
Hess diagram	Plots the relative density of the occurrence of stars at different color-magnitude positions of Hertzsprung-Russell diagram for a given galaxy.
HSB	(High Surface Brightness)
HST	(Hubble Space Telescope)
HTTP	(HyperText Transfer Protocol) An application protocol for distributed, collaborative, hypermedia information systems.
IDL	(Interactive Data Language) A programming language used for data visualization and analysis.
IPAC	(Infrared Processing and Analysis Center)
IRAF	(Image Reduction and Analysis Facility) NOAO image reduction/analysis and visualization software system.
IVOA	(International Virtual Observatory Alliance) The international VO community responsible for developing VO standards.
JIRA	A commercial tool for software teams to plan, build, and track projects.
JPEG	(Joint Photographic Experts Group) Lossy compression for digital images
LDAP	(Lightweight Directory Access Protocol) An industry standard application protocol for accessing and maintaining distributed directory information services over an Internet Protocol (IP) network.
LSB	(Low Surface Brightness)
LMC	(Large Magellanic Clouds)
LSST	(Large Synoptic Survey Telescope)
MAST	(Mikulski Archive for Space Telescopes)
MPC	(Minor Planet Center)
MCs	(Magellanic Clouds)
MySQL	Popular open source database.
MyDB	Simple database wrapper for MySQL.
NASA	(National Aeronautics and Space Administration)
NCSA	(National Center for Supercomputing Applications)
NHPPS	(NOAO High-Performance Pipeline System) An event-driven, multi-process executor system developed to manage pipeline applications in a coarse-grained, distributed processing environment.
NOAO	(National Optical Astronomy Observatory)
NSA	(NOAO Science Archive)
NSSDC	(NOAO System Science and Data Center)
OCD	(Operational Concept Document)
ORD	(Operational Requirements Document)
OS	(Operating System)
PCMS	(Positional Cross-Match Service)
PNG	(Portable Network Graphics) Raster graphics file format that supports lossless data compression.
PSF	(Point Spread Function)
QServ	The LSST database management system.
R	A programming language and software environment for statistical computing and graphics.

RDBMS	(Relational DataBase Management System) A DBMS that represents data using a relational database.
Relational database	A database that stores data in a structure consisting of one or more tables (aka relations) of rows and columns, which may be interconnected.
ReST	(Representational State Transfer) An approach to web services that uses the standard HTTP GET and POST protocols.
SAD	(System Architecture Design) document
SAMP	(Simple Applications Messaging Protocol) A VO protocol for desktop messaging.
SCS	(Simple Cone Search)
SDM	(Science Data Management) group
SDSS	(Sloan Digital Sky Survey)
SED	(Spectral Energy Distribution) Plot of brightness of flux density versus frequency or wavelength.
SExtractor	A program that builds a catalogue of objects from an astronomical image.
SIA/SIAP	(Simple Image Access Protocol) A VO protocol that supports queries for images available in a given data collection near a given position on the sky.
SMASH	(Survey of the MAgellanic Stellar History) PI: Nidever
SMC	(Small Magellanic Cloud)
SN	(Super Nova)
SQL	(Structured Query Language) The standard language used to communicate with RDBMS's.
SQLite	A software library that implement a self-contained, serverless, zero-configuration, transactional SQL database engine.
SRD	(Science Requirements Document)
SSh	Secure Shell
SSA	(Simple Spectral Access) A VO protocol for spectral query/retrieval.
SSO	(Single Sign-On)
SUC	(Science Use Case) document
SVC	An abbreviation for a Web service.
SWIG	(Simplified Wrapper and Interface Generator) An open source software tool used to connect C or C# programs or libraries with scripting languages.
TAP	(Table Access Protocol) A VO protocol for querying remote databases.
TCP	(Transmission Control Protocol) One of the core protocols of the Internet protocol suite, commonly referred to as TCP/IP.
TSV	(Tab-Separated Values) A simple file format often used to move tabular data between computer programs that support the format, e.g., transferring information from a database program to a spreadsheet.
URI	(Uniform Resource Identifier) An address standard for a resource available on the Internet.
URL	(Uniform Resource Locator) The global address of documents and other Resources on the World Wide Web. The address contains 2 parts: specification of the protocol to be used in accessing the resource and its network location.
UWS	(Universal Worker Service) pattern defines how to manage asynchronous execution of jobs on a service.
VAO	(Virtual Astronomical Observatory) The US VO project.
VM	(Virtual Machine)
VO	(Virtual Observatory)

VOSI	(VO Support Interfaces) The minimum interface that a SOAP or REST-based web service requires for compatibility with the IVOA.
VOSpace	The IVOA interface to distributed storage that specifies how VO agents and applications can use network attached data stores to persist and exchange data in a standard way.
XML	(eXtensible Markup Language)
2MASS-PSC	(2 Micron All Sky Survey – Point Source Catalog)

Appendix II: Mandatory ObsCore Data Model Components

Column Name	*Unit*	*Type*	*Description*
dataproduct_type	unitless	string	Logical data product type (image etc.)
calib_level	unitless	enum integer	Calibration level {0, 1, 2, 3}
obs_collection	unitless	string	Name of the data collection
obs_id	unitless	string	Observation ID
obs_publisher_did	unitless	string	Dataset identifier given by the publisher
access_url	unitless	string	URL used to access (download) dataset
access_format	unitless	string	File content format (see in App. **Error! Reference source not found.**)
access_estsize	kbyte	integer	Estimated size of dataset in kilo bytes
target_name	unitless	string	Astronomical object observed, if any
s_ra	deg	double	Central right ascension, ICRS
s_dec	deg	double	Central declination, ICRS
s_fov	deg	double	Diameter (bounds) of the covered region
s_region	unitless	AstroCoordArea	Region covered as specified in STC or ADQL
s_resolution	arcsec	float	Spatial resolution of data as FWHM
t_min	d	double	Start time in MJD
t_max	d	double	Stop time in MJD
t_exptime	s	float	Total exposure time
t_resolution	s	float	Temporal resolution FWHM
em_min	m	double	Start in spectral coordinates
em_max	m	double	Stop in spectral coordinates
em_res_power	unitless	double	Spectral resolving power
o_ucd	unitless	string	UCD of observable (e.g., phot.flux.density)
pol_states	unitless	string	List of polarization states or NULL if not applicable
facility_name	unitless	string	Name of the facility used for this observation
instrument_name	unitless	string	Name of the instrument used for this observation