Sixth Review of the DES Science Portal
4 & 5 August 2014, Fermilab

Process

The Sixth Review of the Dark Energy Survey Science Portal was held at Fermilab August 4 and 5, 2014. The focus of this review was on the Data Server; later (October or November 2014) we will consider how the Portal complements and connects to facilities at the National Center for Supercomputing Applications (NCSA).

We heard informative presentations on Introduction (L. da Costa), Data Server Overview; Installing and Validating Data Releases (A. Fausti), VAC creation, validation, and distribution; Validating external catalogs (R. Ogando), and Data Mining (A. Fausti). Besides the presentations, several documents were placed on the docDB site for this review, with authors including Patricia Egeland, Aurelio Carnero, and Julia Gschwend.

The review was announced to the Science Committee with an open invitation to anyone interested in participating. In addition, targeted invitations were sent to individuals working on topics that were especially relevant to the items in the Charge (included here are Appendix A). The reviewers were, by definition, those people who participated, namely:

Gary Bernstein, Ofer Lahav, Diego Capozzi, Rich Kron, Wyatt Merritt, Eric Neilsen, Stephen Kent, David Finley, Nikolay Kuropatkin, Tom Diehl, Flavia Sobreira, Alex Drlica-Wagner, Jim Annis, Brian Yanny, Huan Lin, and Douglas Tucker. The review was chaired by Rich Kron. Extensive notes from Brian Yanny are included as part of this report.

The format of the review was to hear the presentations on Monday morning, August 4, with an open discussion of issues in the afternoon. On Tuesday August 5 we discussed the set of topics in Appendix B.

The content of this report is largely technical, but we note two aspects of the capabilities of the Portal that relate to the scientific activities of the DES Collaboration. First, it would be useful to have a list of specific science projects which are especially adapted to the Portal. This list could be produced cooperatively between the Science Committee and the Portal development team. Second, DES is launching an effort to understand how best to evaluate catalogs that are derived from the data products produced by NCSA. The Science Portal can play a central role in this process and it will be productive to plan for this activity.

Executive Summary

The portal continues to steadily improve and offer more features that the Collaboration has requested and which are not provided in any other facility within DES. Remarkable progress has been achieved in the past six months, especially regarding the installation of the Portal at Fermilab, which is operating as the Data Server. The Data Server allows full DES-wide Collaboration access to the released products, enabling visualization of the data, plus many other features including tools for data validation. The Portal provides extensive capabilities to create catalogs that are specialized for particular scientific applications. These capabilities include producing catalogs with a subset of source parameters; producing catalogs that include source parameters produced outside of DESDM; matching with wholly external catalogs; and combinations of all of these. These tools are clearly aligned with the needs of the Science Working Groups and we look forward to accelerated usage as more data are ingested and as the DES Collaboration becomes more familiar with the interface and tools.

Significant Points of Discussion

1) Up to now we have mostly considered the properties of the DES on a per-object basis, that is, SQL is run to identify objects that fulfill certain criteria, including data-quality criteria. However, considering the survey as a map of the sky, the quality is more naturally considered as a function of direction in the sky (ra, dec), according to which exposures contributed to the map at that point, and what their individual qualities are. It would be natural to use the DES survey data to ask questions of the map, for example, "what sources are near (ra, dec), what are their properties, and what is the quality of the map there?" A needed facility is to filter a catalog (including a random catalog) based on values in one or more maps.

Recommendation: To the extent it is practical to do so, it would be worth exploring how to connect the data-quality information (say, as considered to be maps) to the source catalogs.

2) An important question to be able to answer is what is the effective survey footprint as a function of data quality. Specifically, if an object were to have properties X, Y, and Z, then within what area of sky could that object have been found? The Mangle masks are used to obtain the depth of the survey at the positions of individual sources, but as mentioned above more flexibility is needed. The Mangle masks provide complete spatial information, but in practice we can usually tolerate finite resolution, say of order 20 - 30 arcsec, as long as the coverage fraction is tracked. See

https://cdcvs.fnal.gov/redmine/projects/des-sci-verification/wiki/Sub-Pixelizing_Mangle_Masks

https://cdcvs.fnal.gov/redmine/projects/des-sci-verification/wiki/Testing_Correlation_Functions_with_Pixelized_Mangle_Masks

Recommendation: Consider using pixelized masks to create depth and systematics maps.

3) The importance of the single-epoch exposures and catalogs came up several times. For example, very often one sees a feature in a co-add tile, and one wants to inspect all of the single-epoch images that went into the co-add at that position. An example at the catalog level is to tag each measurement of flux with the time of the observation to enable a light curve to be produced. Concerning visualization of the single-epoch images, a zoom capability from the whole focal plane to the scale of pixels is desirable, but this is computationally heavy because of creating the png's. Weak lensing needs updated single-epoch astrometric information to facilitate the iteration SE -> coadd -> SE -> coadd.

Recommendation: Continue to explore ways to connect with DES DM database to retrieve single-epoch information.

4) The effort on developing the SQL interfaces is looking very promising. By analogy with SDSS CASJobs, this capability will be a workhorse for Collaboration interaction with DES data. The SDSS experience suggests that most of the time, scientists use simple queries that require correspondingly simple tools. Examples of such queries are: "I have a list of 200 source positions - what is in the DES at (or near) those positions?" "What single-epoch images went into the coadd at a particular (ra, dec)?" "I would like to open the Tile Viewer centered at a particular (ra, dec) and showing a certain field of view, overlay the cataloged DES sources, be able to click on a source to obtain its measured parameters, and be able to download these data."   Development of a suite of relatively simple tools could have a disproportionate benefit in terms of common usage of the Portal.

Recommendation: Continue implementing features in the Fermilab installation of the Portal, updating the portal code and database to include all planned functionality (for example finish the VAC upload tool). Finish the prematching against external catalogs and storage in the database during release installation.

Recommendation: Despite the uncertainty in how the performance of the Data Server scales with number of users, it would be good to invite people in to the Portal at Fermilab now in order to test.

Additional Findings and Comments

Finding: data transfer rate from NCSA to Fermilab is 150 Mbps; this needs improvement. Help from the network experts at Fermilab may be required to diagnose the problem.

Comment: use hexbin in the Tile Viewer because the area will eventually be so large that individual tiles will not be visible on a display.

Comment: concerning the Acceptance Tests, the FWHM requirement is on individual exposures, not on the coadd.

Comment: "Blacklist" has a certain meaning, maybe not the same as what can be done to flag images in the Portal.

Finding: ingestion of external catalogs requires attention to formats. Much hands-on effort is required in the ingestion (not just formats, but also the QA) and perhaps this is inevitable.

Comment: if two catalogs are matched, one should be able to specify whether the result is the intersection or the union of the two lists.

Finding: It will be helpful to develop metrics for the performance of the database: for example, what is needed to serve the DES community in terms of latency? How does the latency and other metrics scale with the number of users?

Comment: concerning the SLR QA tool, it would be useful to have the local value of the reddening apparent also.

Finding: Mangle delivers a molygon table and a coadd molygon table. Some of these products are too large to move around, but that is an argument for making it in one place and using it there.

Comment: Segmentation maps are a product of SExtractor and they tell which pixels are associated with which object. This is important information (an example use case is strong lenses), but the Portal is not set up to handle pixels. A possibility would be to make the segmentation maps look like a bitplane mask. Then, one could overlay the segmentation maps in the Tile Viewer. We may want to have two maps: one for noise and one for the width of the point-spread-function (and perhaps also a map of sky transparency).

Finding: The task of the Spectroscopic Task Force and the photo-z SWG is to create catalogs of redshifts. The task of the Portal is to create training sets and files in a systematic way. The Brazil team should suggest file format and other standards to receive Magellan and Gemini spectroscopic data.

Notes from Brian Yanny

1. User Query feature

A highlight of this review was the demonstration of the 'User Query feature' which allows SDSS-CASJOBS functionality, namely to build and edit an arbitrary SQL query to select subsets of objects from the joint database catalogs present in the Portal, including DES COADD catalogs, external catalogs such as SDSS and 2MASS, and value-added catalogs (VACs) such as tables of shear measures or photo-z's. These user SQL queries are stored and may be edited and resubmitted by the user with custom changes if desired. They may also be shared with other users. The results of the queries are stored in the Portal and can be downloaded by the user. The query results may be turned into their own, documented VACs for long-term storage (such as providing a provenance lineage supporting a data reference or for a figure in a published paper).

Comments on User Query feature:

A. Meets a key need of the collaboration to have a SDSS-CASJOBS like interface to the released data.

B. We would like to have all the User Query features of the main DES science portal replicated in the FNAL copy of the portal.

C. The 'download user query results' feature results in the outputs being divided by COADD tile. If one queried a few objects (such as objects in a narrow color/magnitude range) over a large footprint, this could lead to hundreds or thousands of small output files, each with just one or two objects in them. It would be useful to have an option which merges all the outputs from a query into a single downloadable tarball or ascii csv or FITS binary table file.

D. A user should be able to submit a 'top 10' or 'rownum < 10' version of their SQL query and see the results immediately (within a few seconds to no more than one minute) displayed on a screen without needing to click through many pages to get to a download output screen.

2. Storage and access to maps and masks

We discussed in detail another type of data, orthogonal to object catalog data, namely data that are stored in a 'map' type structure. Examples are the reddening map of SFD98; Planck; and the MANGLE depth-of-field information. A data release's footprint on the sky is a specialized form of a MANGLE map.

The Portal would greatly enhance its usefulness to DES if it could support such map-type data structures, along with fast access mechanisms into these structures. One should be able to efficiently (ideally many hundreds of times per second, depending on the map) probe the map at an arbitrary (ra,dec) position and return the value at this position (possibly interpolated).

A simple example of this is the reddening map of SFD98 or the Planck version of a reddening/dust emission map.

Requirements-by-example on a map support feature.

1. Given an arbitrary (l,b) (or RA, DEC converted to l,b) a function should exist connected to the database query engine such that ebv(ra,dec) would return the E(B-V) value at that position from the pixelized version of the map (or by the four nearest pixels to a given (ra,dec) and interpolating and returning an average E(B-V) value at that point).

2. This function should be linked to the database SQL queries so that one could extract a dereddened magnitude for a coadd object with a query such as:

```
select ra,dec,mag_psf_g, mag_psf_g-3.7*ebv(ra,dec) as g0 from
Y1P1_COADD where ra between 350 and 351 and dec between -0.7 and -0.6
```

3. A more complicated version would be a heal-pixelized version(s) of the MANGLE depth map. For speed, one could store a high resolution pixelized version of the full map, perhaps n=8192 (the size of maps is 8*n^2*wordsize or 3.2GB for a n=8192 map with resolution ~ 0.5 arcmin and a 4-byte stored wordsize, such as depth-in-magnitude).

Then, for each galaxy (or point on the sky), one could determine the magnitude of the galaxy and the depth at that point with a query such as:

```
   select ra,dec,mag_model_g, mag_model_g-3.7*ebv(ra,dec),
   manglemap(ra,dec,'depth-g') as depth from
Y1P1_COADD where ra between 350 and 351 and dec between -0.7 and -0.6
```

Here the 'depth-g' field in the manglemap(ra,dec,whatQuantity) would return the limiting magnitude in the g-band, a number such as 24.20.

One could then also construct a random catalog by probing the manglemap(ra,dec,'depth-g') at a million or a billion random (ra,dec) points in order to construct a data-random, data-data, random-random correlation plot, such as the large scale structure team uses.

4. Performance is important here, it should be able to return hundreds of such manglemap(ra,dec,'depth-g') values per second. The healPy utilities available for python may be useful here.

5. There are also Hierarchical Triangular Mesh (HTM) and SphericalGeometry packages available which perform similar functions to Healpix maps, and these may be considered as well (HTM maps consist of overlapping circles on the sphere which map precisely onto MANGLE polygon boundaries, unlike Healpix pixels which are optimized for C_l spherical harmonic analysis).

3. Support for single-epoch data releases and time series data

The portal now is optimized for ingesting coadd images and tables that come in a tagged data release.

It should be possible, and would be desirable, to also support single-epoch images and tables, provided they also come in some sort of formal tagged release from NCSA/DESDM.

An extremely useful service that the portal could provide would be time series connections between coadd objects and all the single-epoch objects that went into a coadd (a variable length list or time series).

Doing the time series match on (ra,dec), a one-time operation, in the portal would save researchers enormous amounts of time in hunting down single-epoch outliers (data affected by artifacts) as well as greatly facilitating the search for transient, variable or moving objects.

4. Visualization performance and benchmarking

The visualization tools remain a key feature of the Portal. A valuable and well-implemented tool in the current version of the Portal allows quick access to marked pictures of objects associated with any class of query result.

The performance and response time of the Portal to multiple interactive users, either browsing the images and .pngs or running short interactive queries (rather than batch-mode queries) should have some requirements set (i.e. 5 simultaneous users should able to browse a color .png of a tile, panning back and forth, in and out, without noticeable degradation of performance), and these requirements should be benchmarked.

Other notes:

1. The US Department of Energy and the National Science Foundation are moving to a model where, for publications funded by these agencies, it must somehow be possible to reproduce the data used to generate tables and figures. The DES Science Portal would serve as an excellent way to store datasets for the long term used in the construction of such Figures and Tables, and help researchers meet this requirement.

2. We note that catalog uploading in bulk support mode is also improved with the GAVO DaCHS tool.

3. We welcome the use of the VO 'unified content descriptor' tools from the VO to allow the import of a wide variety of heterogeneous catalogs and to further allow the cross match and cross query of these catalogs with SQL queries joined to the coadd_objects and other DES catalogs.

4. We were very pleased to see that the VHS source catalog was able to be ingested into the portal.

5. We were very pleased to see that the shear catalog imshear3 table of shears for all galaxies was able to be ingested into the portal and cross matched to the coadd_objects catalog.

6. We appreciate the versioning system and tagging mechanisms presented for user queries and VACs.

7. We welcome the use of the EEUPS system which also serves the as the DESDM software build and distribution system.

8. We liked the demo videos; they are useful tools to help bring users up to speed on how to use the interface.

Appendix A

DES Science Portal Review: Fermilab, 4 and 5 August 2014

The DES Science Portal, a web-based integrated framework, was designed to enable collaborators and eventually the general public to access a large variety of services and data in a single place, such as:

1. Results of the validation of the DECam data carried out at the telescope (Quick

Reduce)
2. Evaluation of the results of single-epoch exposures (in progress)
3. Validation of the coadds (QA)
4. Visualization of images and catalogs
5. Production of value-added catalogs with photo-z, galaxy properties, star/galaxy separation and sample selection with full provenance and characterization
6. Distribution of DES value-added catalogs produced by the portal or by other authors with full characterization and provenance.
7. Distribution of ancillary files (training sets, mangle, systematic maps, depth maps, spectroscopic data)
8. Access to simulations (BCC, MICE)
9. Access to other sky surveys (APASS, NOMAD, UCAC3. 2MASS, WISE)
10. Combination of DES catalogs with other surveys (VHS, WISE, SPT)
11. Access to a code repository
12. Availability of science workflows that can be used on demand for catalog validation and/or for key project analysis

Goal for Monday is to ascertain the following:

1. The infrastructure needed for data validation and catalog production
2. What is needed for the portal to serve as an effective user interface to the DES data at NCSA
3. the best way to use the Science Portal to characterize catalogs produced by the collaboration
5. The operational responsibilities of the LIneA team and its operational readiness
6. How best to connect to resources at Fermilab and NCSA

In the process of evaluating the above list of functionalities, the following topics should be discussed:

- hosting the portal: computational resources and connectivity need to be fast and serve a large number of collaborators
- review the functionalities available in the Tile Viewer
- define how, for what, and by whom VACS will be generated
- how the portal can coalesce the effort being carried out by a number of individuals and sub-groups

Appendix B

Topics for discussion Tuesday, August 5, 2014

1. What are the ancillary files needed for a "complete" Release and how are they created?

Mangle

depth map
systematic map
random catalog
photo-z training set
photo-z training files
photometric calibration (SLR, extinction correction)

2. Given a co-add Release, what are main "added values?"

which photo-z's codes?
which S/G codes?
what galaxy properties?

3. What are the selection criteria needed to prepare a value-added catalog that is ready for analysis?

which flags?
S/N or mag limit?
discard non-entries (e.g. -99)?

4. What is needed in order to select by region in the sky (as opposed to selecting by object criteria)?

5. What tests are needed to characterize VAC's (context is standardization to help the Collaboration)?

6. Other topics if not already covered

image viewer to examine catalog and target lists
centralized spectroscopic database
set of sample queries
desired work priorities for Brazil