LSST DATA CHALLENGE HANDBOOK

HANDBOOK FOR THE ANALYSIS OF LSST DATA AND CATALOGS

User Support

There are multiple sources of help for accessing or understanding LSST Data Challenge products and services. Please see *Chapter 1.4.* Obtaining Help for details.

User Forum: https://www.lsstcorp.org/sciencewiki/index.php?title=Special:AWCforum

Users may also find the Science Collaboration Wiki to be helpful.

Science Wiki: http://www.lsstcorp.org/sciencewiki/index.php?title=Main_Page

For help with understanding details of the processing algorithms, or with problems with data access, please e-mail the Data Management Help Desk.

E-mail: dc-help@lsst.org

For web-based access to data, see:

Image access: https://osiris.ipac.caltech.edu/cgi-bin/LSST/nph-lsst

Catalog access: https://osiris.ipac.caltech.edu/

Data Quality products: http://lsst1.ncsa.uiuc.edu/pipeQA/public/

DC3b/PT database schema: http://lsst1.ncsa.uiuc.edu/schema/index.php?sVer=PT1_2

Revision History

Version	Date	Editors
1.0	2011 January	Richard A. Shaw, Michael A. Strauss
1.1	2011 August	Richard A. Shaw, Michael A. Strauss

This *LSST Data Challenge Handbook* is available on the Science Wiki (which requires a login):

Handbook: https://www.lsstcorp.org/sciencewiki/images/DC_Handbook_v1.1.pdf

Citations to this *Handbook* should read:

Shaw, R. A., & Strauss, M. A., ed. 2011, *LSST Data Challenge Handbook* (Version 1.1; Tucson, AZ: LSST Corp.)

The front cover features an image of the LSST mirrors, with people for scale, created from mechanical drawings by Todd Mason, Mason Productions, Inc.

LSST is a public-private partnership. Design and development activity is supported in part by the National Science Foundation. Additional funding comes from private foundation gifts, grants to universities, and in-kind support of Department of Energy laboratories and other LSST Member Institutions. The project is overseen by the LSST Corporation, a non-profit 501(c)3 corporation formed in 2003, with headquarters in Tucson, AZ.

Copyright © 2011, LSST Corporation. All rights reserved.

Table of Contents

Preface

Scope of This Document

The purpose of this *Handbook* is to describe the LSST Data Management processing, and the data products it produces, in enough detail that a Science Collaboration member who is not familiar with them can evaluate their quality, scientific fidelity, and suitability to support their science goals with LSST. One of the key aims of the *Handbook* is to provide one-stop shopping for vital background material for the input data and the processing algorithms, with references to more detailed information where it exists. However, this *Handbook* is not intended to replace requirements documents, technical specifications for hardware, or software design documents. It is also not a user guide for data analysis, although advice is offered for some software that may be useful. An excellent overview of the LSST mission and technical capabilities can be found in the *LSST Science Book*, which is available at http://adsabs.harvard.edu/abs/2009arXiv0912.0201L. This document describes the data challenges in general and the accomplishments to date (see Chapter 1), the content and packaging of the data products (images, catalogs, and ancillary files) and how to access them (Chapter 2), the input raw data and the telescope/camera system that was used or simulated to obtain them (Chapter 3), the key algorithms and the processing flow that were used to produce the data products (Chapter 4), and an assessment of the scientific quality of the output data (Chapter 5).

Source Material & Attribution

The material for this *Handbook* was drawn from a large number of sources, including LSST technical documents, Wiki pages, external web pages, mail exploders, data file headers, software design documents, and informal conversations with experts. Often, figures, tables, and even text are excerpted from these sources. In order to keep the style and presentation relatively clean, the attribution to the source material is cited the first time it is used in the main body of each chapter. The last section of each chapter is devoted to a listing of references, contributing authors, and background resources that provide details that fall outside the scope of this *Handbook*.

Special Notes

Selected fonts are used to indicate special terminology, or to indicate user interaction with tools:

- ***Bold italic*** indicates technical terms of interest the first time they are used in the text, and are defined in the Glossary at the end of this *Handbook*.
- Names of software tools or packages are indicated with mixed-case **bold**.
- `Fixed-width type` indicates text that should be input to a software application or web tool.
- **Underlined Arial bold in teal** indicates text that appears on a "button" in an application.

The planned processing software for this data challenge has not been fully implemented, and will grow as the data challenge proceeds (see Chapter 1). Descriptions of processes or data products that are planned but not yet implemented are denoted in tables and figures with a grey background.

Special notes appear throughout this *Handbook* to convey information of special interest or urgency. They are the following:

Informative notes are denoted with the light-bulb symbol, and generally contain tips and pointers that deserve special attention.

Cautionary notes are indicated like this, and indicate potential limitations of the data, the instrument, or the processing software that may affect the use or interpretation of the data products.

Warnings of serious consequences are indicated like this, and denote problems with the data that could lead to erroneous scientific or technical interpretations, or problems with software that could lead to errors that may not be apparent to users.

Chapter 1: Introduction to LSST Data Challenges

The LSST Data Management team has been carrying out a series of ***Data Challenges***, ever more sophisticated realizations of the LSST data and the pipelines and infrastructure needed to analyze them. This chapter describes the purpose and plan for the current data challenge, and a rough timeline for the production of the data products. The output data products, including images and catalogs, are described in Chapter 2. The final products are derived from two primary sources of raw data, the general properties of which are described in Chapter 3. The processing flow and applicable algorithms are described in some detail in Chapter 4. This chapter includes a summary of applicable policies for use of these data and for scientific or technical publications that may be produced; it concludes with a description of how to obtain help in accessing or analyzing data from the current data challenge.

1.1. Data Challenge 3b in a Nutshell

The current data challenge, termed *DC3b* (Kantor et al. 2010), uses both real data from the *CFHT Legacy Survey* (***CFHT-LS***) and extensive image simulations of LSST data (***ImSim***), and has as its goal to prototype most portions of the full Data Release ***Production*** system, including the so-called *Multi-fit* algorithm for detecting and measuring the properties of faint objects in multiple repeat visits to a given area of sky. DC3b is being carried out in multiple phases, with corresponding semi-annual data releases.

Each phase of DC3b delivers different types of data and steadily improving levels of data quality. The first phase, released in January 2011, was largely intended for initial software integration and shakedown. The general goals were to remove instrumental signature from and calibrate single-visit images, perform PSF photometry on detected sources, and generate a catalog of objects from multiple detections of sources. The amount of data processed was relatively small, the processing stages are limited, and data quality goals were not uniformly achieved. The input data for the first release included the first large-scale runs of the LSST simulation framework. These simulations were at an early stage of maturity and had not been validated against the expected performance of the LSST hardware, i.e., they were not intended be used to determine the final capabilities of the LSST.

The second phase of the DC3b effort is complete, and access to the data products is now available, as before, with a user interface based on **Gator**, **VOInventory**, and tools developed by the Data Management team. These data are expected to be of markedly greater interest to the science collaborations than the previous release. Single frame measurements should reliably meet data quality requirements, the astrometric solutions and single-visit photometry should be more accurate, as should time series of transient and variable objects.

LSST Data Management encourages the Science Collaboration members to use DC3b data to become familiar with the catalog database and access tools, and to perform quality checks on the data that may uncover problems that have not been documented. However, collaborators should not expect to derive useful science from these data products or to use them for evaluation of their LSST science programs. In particular, the current production software does not include creation of image co-adds or any of the data products that depend on them, nor does it include detection and orbit determination of solar system objects. The simulations are, however, at a stage where the simulation team would welcome feedback on the images and catalogs including comparison with existing deep imaging data sets, and requests for additional capability in the input catalogs and the image simulator.

A future data release, probably in 2012, will add the production of image co-adds for both the creation of difference images and detection and measurement of faint objects, the determination of astrometric models (proper motion and parallax), and asteroid orbits from the moving-object pipeline (MOPS), though these are likely to be restricted to slow movers (i.e., main belt and beyond). Once the production system has met this level of maturity, single frame measurements should reliably meet data quality requirements, the astrometric models derived from them should be valuable, as should time series of transient and variable objects. MOPS results will be preliminary, although shape measurements of faint galaxies may initially be relatively primitive.

The last phase of DC3b is expected to add full-fledged *Multifit* measurements of faint galaxies, and an initial implementation of global photometric calibration for the simulated image data. Results from this data release are planned for 2012, and we expect they will be useful for evaluation of LSST science programs, and possibly even for doing new science with CFHT-LS. Your help with evaluating the data from each release will help ensure that this is the case!

1.2. Goals and Status for the Current Performance Test

The plan for most recent round of processing, released in Summer 2011, included several high-level goals for what data would be processed, and the science quality of the results. These goals and their status are summarized below. Goals that were not achieved for this release have been deferred to the next phase or a later data challenge.

- Generate 449 simulated LSST visits (2 exposures per visit), distributed among 7 adjacent areas of sky (each covering roughly 10 deg^2) in six passbands: *u, g, r, i, z,* and *y*. **Status: achieved** (no *u*-band visits were available for this time period, however).
- Process all visits through the Data Release Production, and produce calibrated, single-visit images plus photometric catalogs of sources and objects. **Status: achieved (failure rate of ~0.05% of CCDs)**.
- Generate automated data quality reports that facilitate the assessment of the accuracy of the astrometric solutions and photometric zero-points. The reports are available at: http://lsst1.ncsa.uiuc.edu/pipeQA/public/ **Status: achieved**.
- Assure that the headers of reduced images are populated with updated keywords to reflect provenance, calibrations, and other operations that have been performed, and that keyword comments are retained. **Status: achieved**.
- Demonstrate that the accuracy goals for astrometric and photometric fidelity have been met:
    - World Coordinate System solution: <200 mas RMS. **Status: not uniformly achieved**. Note, however, that this result does *not* reflect the achievable accuracy of the global astrometric solution.
    - Photometry of isolated point-sources: <0.05 mag RMS. **Status: achieved.**
    - Photometry of small galaxies: <0.07 mag RMS. **Status: not achieved.**
    - Star-galaxy separation—i.e., reasonably reliable flagging of point sources from non-point sources. **Status: achieved**.
- Stretch goal: Produce rough galaxy shape measurements using an early (and not fully functional) version of the *Multifit* pipeline. **Status: not achieved.**

See Chapter 5 for a more detailed discussion of data quality. Generally speaking, this data release provides single-visit calibrated images (without co-addition of paired exposures), basic source and object catalogs, rough star/galaxy separation, and the identification of variable objects. Left for future

releases are deep image stacks, pan-chromatic detection stacks, difference images, the identification of moving objects, multi-object fitting and measurement (for crowded fields and complicated targets), and global astrometric and photometric calibration.

1.3. Publication and Data Use Policy

Results from the analysis of data products from DC3b may appear in internal reports, as well as presentations, external technical publications, and (possibly) science papers. Authors of papers and reports that make use of data products produced by LSST software, or the LSST software components themselves, should be aware that publications should include an appropriate acknowledgement of the source of these resources. In addition, there is a formal policy for publications of LSST Science and Technical material (LSST Science Council, 2011). This LSST Publication Policy provides direction on authorship, attribution, and acknowledgements, and requires an internal review of the content by the relevant Science Collaboration or other authority prior to publication.

For those who make use of CFHT-LS, please be aware of your obligation to include the following acknowledgement in any publication using these data:

> *Based on observations obtained with MegaPrime/MegaCam, a joint project of CFHT and CEA/DAPNIA, at the Canada-France-Hawaii Telescope (CFHT) which is operated by the National Research Council (NRC) of Canada, the Institut National des Science de l'Univers of the Centre National de la Recherche Scientifique (CNRS) of France, and the University of Hawaii. This work is based in part on data products produced at TERAPIX and the Canadian Astronomy Data Centre as part of the Canada-France-Hawaii Telescope Legacy Survey, a collaborative project of NRC and CNRS.*

1.4. Obtaining Help

There are multiple sources of help for accessing or understanding LSST Data Challenge products and services. The primary source of assistance, beyond this *Handbook*, is the Data Challenge User Forum. The intent of the forum is to create a place for LSST users to ask questions of other users. Although the forum will be monitored by LSST staff (who will also answer questions and participate in discussions), the goal is to provide an independent source of support for the LSST community, drawing from the collective experience of its own members. A forum format is well suited for questions and their subsequent discussion and resolution. An e-mail interface to the forum may be provided (for users who prefer to receive an email digest) depending upon demand. In addition, the forum will provide a searchable archive of previously asked questions, which users should consult prior to asking their own question. The ultimate goal is that the forum become a useful and self-sufficient resource for the LSST community, where users may gain insight into common (or rare!) problems and contribute to the growing understanding of LSST, its data characteristics and the data interfaces, while continuing to build a sense of community.

The User Forum can also facilitate the coordination of the scientific analysis and data quality assessment across Science Collaborations, and will include advice on tools, techniques, and avoiding pitfalls. To access the User Forum (which requires a login and password) go to:

https://www.lsstcorp.org/sciencewiki/index.php?title=Special:AWCforum

Users may also find the Science Collaboration Wiki (which requires the same login and password) to be helpful. You can access it via:

http://www.lsstcorp.org/sciencewiki/index.php?title=Main_Page

Finally, for help with understanding details of the processing algorithms, or with problems or technical issues related to data access, or for obtaining a login for the Science Collaboration web site,

please e-mail the Data Management Help Desk. The Help Desk will address questions on a best-effort basis, with a goal of resolving issues within a few business days.

E-mail: dc-help@lsst.org

1.5. References and Further Information

Contributing Authors

Contributors to the technical content of this chapter include Tim Axelrod, Lynne Jones, Jeff Kantor, Dick Shaw, and Michael Strauss.

References

Kantor, J., Axelrod, T., Allsman, R., Freeman, M., Lim, K.-T. 2010, *Data Challenge 3b Overview*, LSST Document 9044 (Tucson: LSST Corp.), available at: http://www.lsstcorp.org/sciencewiki/images/DC3b_Scope.pdf

LSST Science Council 2011, *LSST Publication Policy*, LSST Document 7644 (Tucson: LSSTC), available at: https://www.lsstcorp.org/sciencewiki/images/LSST_publication_policy.pdf

Chapter 2: Accessing LSST Data

The model that has been adopted for user interaction with LSST DC3b data is to provide users with the ability to search for, select, access, and retrieve data products of interest to them, with users analyzing the data on their personal compute platforms. Other models, such as providing compute resources on the LSST cluster, access to the LSST software stack and (easily) configurable reduction schemes, and the use of user-contributed software for bulk processing of LSST data are planned, but are not yet supported. This chapter describes the content and structure of the data products that have been produced for DC3b, as well as the process for searching and retrieving them. Software that may be helpful for data retrieval and analysis is summarized at the end of this chapter.

2.1. Output Data Products

The data products that are produced by the production pipelines consist of images and catalogs, the contents and structure for which are described in the following subsection. The structure and other characteristics of the input raw images are described in Chapter 3.

2.1.1. Catalogs

The catalogs that are populated by the pipelines, listed in Table 2-1, are likely to be more extensively used than other kinds of data products, both for science and data quality evaluation. The output catalogs are stored as tables in the ***Science Database***[1]; the ImSim input catalog of objects is stored in a separate database. It is easiest for users to query and analyze portions of the catalogs using the **Gator** interface, which is described in Section 2.2.1. below.

Table 2-1: Science Catalogs

Catalog Type	Description
Exposure[2]	Describes each exposure, including the date/time of exposure start, the filter used, the position and orientation of the FoV on the sky, and other environmental information.
Source	Describes each detected source on each Calibrated Image (see Sect. 2.1.2.), including its location (*x,y*) on the detector, world coordinates (RA, Decl), brightness, size, and shape.
Object	Describes attributes of each astrophysical object, including the world coordinates, brightness in each color with time, etc.
Moving Object[3]	Attributes of moving (solar system) objects, including orbital elements, brightness, albedo, rotation period.
ImSim Input Objects	Catalogs of objects that were used by ImSim to generate images. Includes object world coordinates, type, size, shape, brightness, and orientation.

It is worth emphasizing the distinction between the terms ***source*** and ***object***. A ***source*** is a detection of an astrophysical ***object*** in a single image (i.e., an exposure), in a single passband, the

[1] The Science Database schema for the June, 2011 Data Release may be browsed at http://lsst1.ncsa.uiuc.edu/schema/index.php?sVer=PT1_2

[2] Of the multiple tables listed in the Science Database schema browser that contain exposure metadata, the content of **Science_Ccd_Exposure** most closely matches the fields described in the public interface.

[3] The Moving Object catalog is planned for DC3b, but not yet available.

characteristics for which are stored in the Source Catalog of the science database. The Data Management System attempts to associate multiple source detections in all passbands to single astronomical ***objects***, such as a star, galaxy, asteroid, or other physical entity, which can be static or change brightness or position with time. Usually an ***object*** will be associated with more than one instance of a ***source*** detection, with the exception of certain classes of transient objects.

2.1.2. Image Products

The content of the image products that are processed by the pipelines (see Chapter 4) are listed in Table 2-2. The files are all in FITS (Pence et al. 2010) format, with very similar but not quite identical internal organization.

Table 2-2: Types of Science Images[4]

Type	Extension Contents	Size	Units	Description
Raw Image	[none]	Amp	ADU	Raw data as obtained from the real or simulated observing environment, formatted as images from individual amplifiers.
Calibrated Image	1: Science	CCD	Electron	Images are corrected for instrument signature; paired exposures are combined with CR-rejection, and background-subtracted; calibrations are determined for WCS and photometric zero-point.
	2: Mask	CCD	[None]	Bit-encoded data quality mask: see Table 2-5 for definitions.
	3: Variance	CCD	$Electron^2$	Variance of Science image, which includes shot noise, read noise, contributions from the noise in calibration reference images, and (for ImSim only) co-addition of the paired visit exposures.
Template Image	MEF: 3	Sky Tile	TBD linear	Result of combining multiple Calibrated Exposures per passband, and removing moving objects and transients
Difference Image	MEF: 3	CCD	TBD linear	Difference between a Calibrated Exposure and warped, scaled Template Image
Deep Co-addition	MEF: 3	Sky Tile	TBD linear	Stacked calibrated science images, one per bandpass, with moving objects removed
Deep Detection Co-addition	MEF: 3	Sky Tile	TBD linear	Stacked, pan-chromatic[5] science image, with moving objects removed

The image sizes depend upon the details of focal plane array of the instrument that generated the data (see Chapter 3 for details), given in Table 2-3 below.

Table 2-3: Sizes of Science Images

Type	Data Source	Image Array Size
Amp (Excluding overscan)	ImSim	509 × 2000 pixel
	CFHT-LS	2048 × 4612 pixel
CCD	ImSim	4072 × 4000 pixel
	CFHT-LS	4096 × 4612 pixel

[4] Rows with a grey background indicate data products that are planned for DC3b, but are not yet being produced.
[5] Pan-chromatic science images will be combined using the algorithm of Szalay, et al. (1999).

Type	Data Source	Image Array Size
Sky Tile	[All]	Varies by sky position. Typically 0°.5 x 0°.5, or 9000 x 9000 pixel

Structure of the FITS Images

The LSST image files differ somewhat in their internal organization, depending on the type of information they contain. They all contain a science array from a single CCD detector, and most products also include pixel-level concomitant data as well: a variance array, and a data quality mask. The basic organization is shown in Figure 2-1 below. The *input* raw images (described in Chapter 3) are organized as simple FITS images—i.e., a header plus science data array in the primary ***Header Data Unit*** (HDU). The *output* images are stored as a primary header plus three ***image extensions***: one each for the science, mask, and variance arrays. In all cases the metadata (i.e., the keyword-value pairs) found in the primary HDU are applicable to all extensions in the file; metadata found in extension headers apply only to that extension.

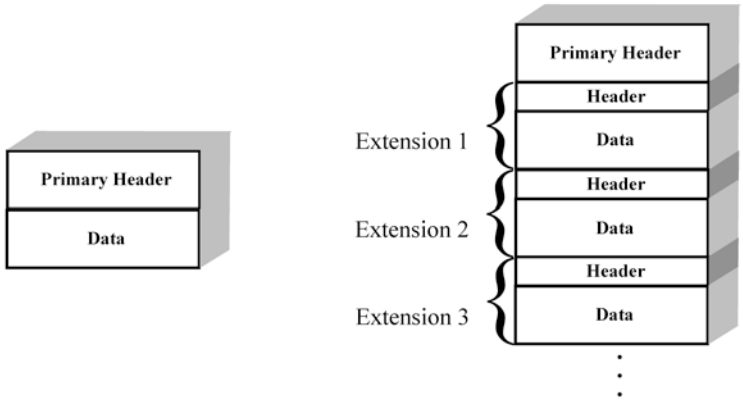

Figure 2-1: Schematic of the structure of a simple FITS file that stores a single image array (*left*) and a Multi-Extension Format (MEF) file that stores multiple components of an image (*right*).

Packaging of the Images

The packaging of image data products for LSST is a compromise among the competing needs of data-parallel processing (which requires small- to medium-sized files), efficient storage (where larger files are optimal), and rapid and reliable transport to users over the internet (for which modest sized, compressed files are optimal). For DC3b, users will have access to images both individually, and in some cases as aggregated into Unix tar files. Raw images are compressed with **gzip**.

The strategy for packaging images is also affected by the way in which the raw data were received from the observing environment (which is different for the CFHT-LS and ImSim data: see Chapter 3), and the scheme for tagging observations—i.e., how data files are named. Understanding the tree-based organization and file nomenclature is key to understanding how one image relates to another. The file/path naming conventions for the various image products is given in Table 2-4, and the fully qualified name is the concatenation of the base path, path and filename.

Table 2-4: Image File Naming Convention

Image Type	Base Path	Path/Filename
	ImSim	
Raw	**/ImSim/raw**	**/v**[*visit*]**-f**[*filter*]**/E**[*exp*]**/R**[*raft*]**/S**[*sensor*]**/**[*fname*]**.fits.gz** **/v**[*visit*]**-f**[*filter*]**/E**[*exp*]**.tar**
Bias	**/ImSim**	**/bias/v0/R**[*raft*]**/S**[*sensor*]**/**[*fname*]**.fits.gz** **/bias/v0/R**[*raft*]**.tar**
Dark		**/dark/v1/R**[*raft*]**/S**[*sensor*]**/**[*fname*]**.fits.gz** **/dark/v1/R**[*raft*]**.tar**
Flat		**/flat/v2-f**[*filter*]**/R**[*raft*]**/S**[*sensor*]**/**[*fname*]**.fits.gz** **/flat/v2-f**[*filter*]**/R**[*raft*]**.tar**
Calibrated	**/ImSim/calexp**	**/v**[*visit*]**-f**[*filter*]**/R**[*raft*]**/S**[*sensor*]**.fits** **/v**[*visit*]**-f**[*filter*]**/R**[*raft*]**.tar**
	CFHT-LS	
Raw	**/CFHTLS/**[*field*]**/raw**	**/v**[*visit*]**-f**[*filter*]**/S00/c**[*ccd*]**-a**[*amp*]**.fits.gz** **/v**[*visit*]**-f**[*filter*]**/S00.tar**
Bias	**/CFHTLS/calib**	**/bias/v**[*visit*]**-f**[*filter*]**/R**[*raft*]**/S**[*sensor*]**.fits.gz** **/bias/v**[*runID*]**-f**[*filter*]**.tar**
Flat		**/flat/v**[*runID*]**-f**[*filter*]**/c**[*ccd*]**-a**[*amp*]**.fits.gz** **/flat/v**[*runID*]**-f**[*filter*]**.tar**
Fringe		**/fringe/v**[*runID*]**-f**[*filter*]**/c**[*ccd*]**-a**[*amp*]**.fits.gz** **/fringe/v**[*runID*]**-f**[*filter*]**.tar**
Calibrated	**/CFHTLS**	**/**[*field*]**/calexp/v**[*visit*]**-f**[*filter*]**/c**[*ccd*]**.fits**

In the above Table, bracketed words in italic are identifiers, and characters in boldface are literal. The *visit* is an integer that identifies the visit, and the single-character *filter* identifies the filter that was deployed during the exposure (one of *u, g, r, i, z*, and for ImSim, *y*). The *raft* and *sensor* refer to the tagging of CCD detectors in the focal plane array, as described in Chapter 3. The CCDs (or ***sensors*** in the engineering vernacular) in the LSST camera (see Figure 3-1) are arranged 3×3 on ***raft*** structures, and labeled *S00* through *S22*. The rafts are organized in a regular grid, and labeled R01 through R43.

For CFHT-LS data, the visit identifier is identical to the running exposure number for MegaCam, as it would be found in the CADC archive. The CCDs are arranged in a rectangular grid (see Figure 3-4), and labeled *c01* through *c35*; there is no equivalent to the raft structure. Also for CFHT-LS, the images are divided by the survey *field* (one of *D1—D4* or *W1—W4*), as described in Table 3-4 on page 22.

Raw data are collected at the level of individual amplifiers (or ***channels***, in the engineering vernacular), and the ImSim files have some additional structure compared to processed images. Specifically, the *fname* incorporates some of the path information:

imsim_[*visit*]**_R**[*raft*]**_S**[*sensor*]_C[*channel*]_E[*exposure*]**.fits**

Some of the individual files are compressed using the **gzip** program (use **gunzip** to uncompress). Raw images for CFHT-LS are not processed as paired exposures (akin to the LSST "visit" of two consecutive exposures at the same pointing), so the *exp* is omitted from the rule above. Finally, while modest file sizes are more practical for transport and real-time science analysis than an entire focal

plane, they can be inefficient when analyzing extended objects, or performing data quality analysis over the full FPA. For this reason, most data can also be retrieved bundled in a Unix *tar* file.

Here are examples of fully qualified path/filenames. The first is for a calibrated ImSim *r*-band image from sensor S02 on raft R01 for visit 85408535:

```
/ImSim/calexp/v85408535-fr/R01/S00.fits
```

The second example is for an entire focal plane of CFHT-LS raw, *r*-band images for exposure 695854, bundled into a tar file:

```
/CFHTLS/D3/raw/v695854-fr/s00.tar
```

CFHT-LS data have yet not been processed at productions scale and are not available for analysis. These data products are planned for a later data release.

Pixel-Level Concomitant Data

Masks

The mask image (extension 2) flags the various pathologies and other attributes of pixels in the science image; their meanings are given in Table 2-5. Each bit has a true (set) or false (unset) state. Flagged conditions correspond to specific bits in a 16-bit integer word. For a single pixel, this allows for up to 15 data quality conditions to be flagged simultaneously (thus far only 7 bits are defined), using a bitwise logical OR operation. Setting none of the bits, or a value of zero in the mask, indicates the pixel is suitable for science use and that no other special conditions apply. (But note that bits 5 and 6, when set, merely indicate the detection of a source, rather than compromised science quality). Note that the data quality flags cannot be interpreted simply as integers but must be converted to base-2 and interpreted as flags. These flags are set and used during the course of processing, and may likewise be interpreted and used by downstream pipeline stages or analysis applications.

Table 2-5: Meanings of Image Data Quality Mask Bits[6]

Decimal Value	Hex Value	Quality Condition Indicated
1	0x**1**	Static bad pixel (e.g., bad column, charge trap)
2	0x**2**	Saturated bad pixel
4	0x**4**	Pixel flagged for interpolation in the science array
8	0x**8**	Pixel compromised by cosmic ray
16	0x**10**	Pixel in the edge region of a detector array, which is the half-width of the smoothing filter used for source detection, typically ~10 pixels
32	0x**20**	Pixel lies within the ***footprint*** of a detected astrophysical source
64	0x**40**	Pixel lies within the ***footprint*** of a detected source in a Difference Image; in this case the source is dimmer than its counterpart in the Template Image, resulting in a negative brightness profile.

Variance Arrays

The variance array describes the statistical uncertainty of the Science array at the pixel level. This is necessary because the processing for any given pixel involves many factors, including data quality

[6] The assignment of named conditions to particular bits is subject to change, but probably not during DC3b.

bits that may be set, pixel-level operations with other images that themselves have variance arrays, and the creation of image stacks whose component images likely do not align perfectly. The variance of the input raw images is estimated from a Poisson model.

2.1.3. Calibration Reference Data

As described in Chapter 4, calibration reference data are used to remove instrumental signature from the raw science frames, and to provide the basis for astrometric and photometric calibration. The image data consist of the products named in Table 2-6.

Table 2-6: Types of Calibration Reference Data

Type	Structure	Size	Description
Bias	MEF: 1	Amp	Corrects residual bias structure that remains after overscan correction
Dark	MEF: 1	Amp	Corrects for dark current, scaled to exposure duration. The dark current in CFHT-LS images is extremely low, so the dark correction is not applied for these data.
Flat	MEF: 1	CCD	Corrects for pixel-to-pixel photometric non-uniformity, and for camera vignetting across individual CCDs.
Fringe	MEF: 1	CCD	Corrects for fringing of atmospheric emission (*i*-, *z*-, and *y*-bands only)
PSF	MEF: 1	CCD	Derived PSF shape from field stars

2.2. Browsing, Queries, and Retrieval

2.2.1. Catalog Data

Public access to the output catalogs is provided through a web interface, shown in Figure 2-2, which is based on the **Gator** catalog tool developed at IPAC. This is a separate database from that used by the Data Management development team, and it has slightly less content—e.g., it lack certain data quality measurements. This database and the **Gator** tool are meant to provide a simple access mechanism to the catalog data even by users who are not conversant with the standard query language, SQL. While **Gator** translates user input into SQL behind the scenes, it does not provide users with the full functionality of SQL, including the ability to perform joins among database tables. To work around this limitation, some joins have been performed in advance, and are available in the interface as "separate" catalogs. Users should select a catalog of interest by clicking the appropriate link on the catalog summary page, which may be found at https://osiris.ipac.caltech.edu/. Please note that this page requires login information, which for members of the Science Collaborations is the combination "`lsst`" and "`Big-Sky`".

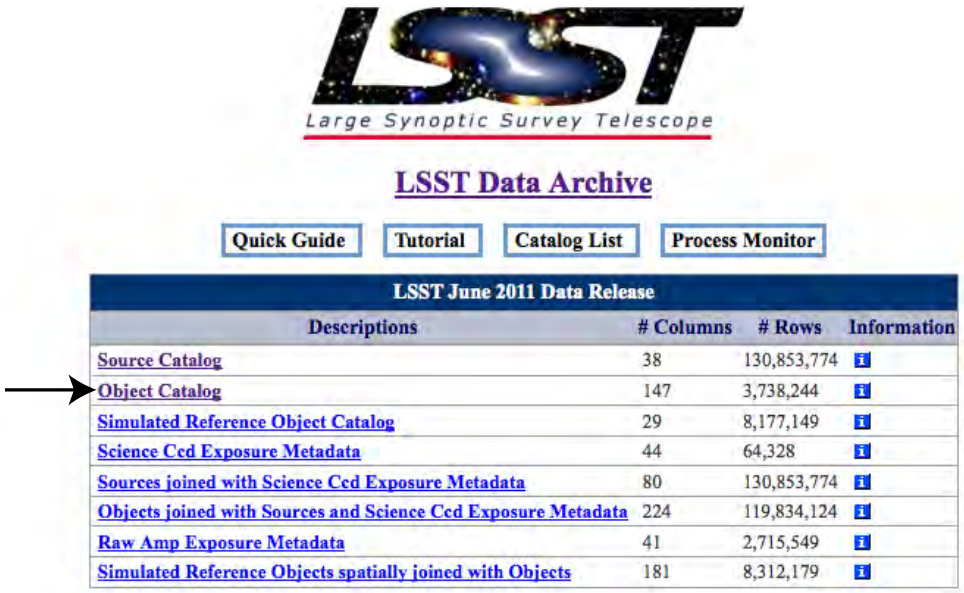

Figure 2-2: Web interface for LSST catalogs that were generated with DC3b processing pipelines.

The selection of catalog entries takes place in response to a user query, which is constructed similarly for all of the catalogs. A search of the Object Catalog for ImSim data involves steps like the following:

1. Click on the name of the desired catalog (ImSim *Object Catalog* in this example).
2. Click the radio-button for "All Sky Search" or enter sky coordinates (e.g., "0 0") plus the size of the search area.
3. Choose which fields will be reported in the output table by selecting the checkboxes in the **Sel** column. (In some cases it may be most efficient to first click the **Clear All Selections** button and then re-select the desired fields by checking the desired field names.)
4. Enter restrictions on the field contents by entering expressions in the "`Low Limit`" or "`Up Limit`" columns. In this case, *r*- and *i*-band object fluxes are restricted to positive values by entering "`>0.`" in the `Low Limit` boxes for the fields **rFlux_PS** and **iFlux_PS**.
5. At the bottom of the entry form is an optional text box for entering additional constraints on the field values or relationships between them, with an SQL-like syntax. In this case, to select objects that have a somewhat blue color, enter the text "`rFlux_PS > iFlux_PS`".
6. Click the **Run Query** button. The constraints will be summarized while the query is executing, which in this case is the expression:

   ```
   (rFlux_PS > iFlux_PS) and rFlux_PS >0. and iFlux_PS >0.
   ```

   The time elapsed for the query will be indicated. It is possible to run queries in background, and be notified via e-mail once the results are available.
7. Once the query has completed (which may take several seconds to minutes), a map indicating the object positions will be displayed, along with a subset (nominally the first 100 rows) of the output table. Hyperlinks are available to view or download the full table.

Links to help for the Gator interface are given at the top of the query page. It is worthwhile to read either the Quick Guide[7] or the short Tutorial[8]. It may also be helpful to explore the LSST Database schema browser, which describes in detail each field in every table of the science database.

Support in DC3b for direct SQL queries of the Science database, and for programmatic access and storage of intermediate results, is under development. In the mean time, experienced database users with advanced query needs, meaning those queries that are not supported with the **Gator** tool, should contact the DM Help Desk at dc-help@lsstcorp.org.

2.2.2. Images

Method 1: VO Inventory Interface

Public access to the raw and processed images is provided through a web service, the interface for which can be found at https://osiris.ipac.caltech.edu/cgi-bin/LSST/nph-lsst (login information is the same as that for catalog access through **Gator**). Figure 2-2 shows the interface, which consists of text-boxes for specifying the search parameters, choices for image collections, and direct links to images. The initial interface (before starting a search) is what appears above the dashed line in the figure.

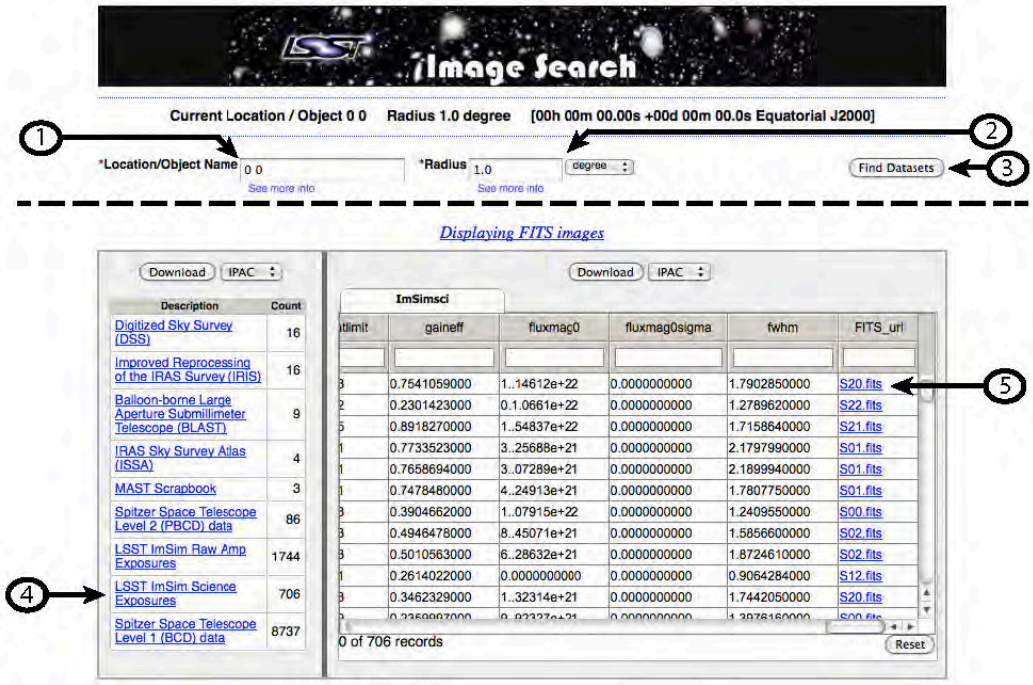

Figure 2-3: Inventory interface for image searches. Numbered callouts are superimposed, which correspond to a sequence in the search process that is explained in the text.

[7] See the Quick Guide to Gator at https://osiris.ipac.caltech.edu/applications/Gator/GatorAid/lsst/quick.html
[8] See the brief Gator Tutorial at https://osiris.ipac.caltech.edu/applications/Gator/GatorAid/lsst/tutorial.html

As the numbered keys in Figure 2-2 indicate, the process for identifying and retrieving an image is the following:

1. Enter the coordinates of an LSST field of interest, or alternatively the name of an astrophysical object that is known to be in the field of images of interest. Clicking the highlighted text just below this text box will show examples of acceptable format. The locations of the sky fields that were simulated are given in Chapter 3, in Table 3-2 on page 19, or that were observed for CFHT-LS in Table 3-4 on page 22. In this example, coordinates of RA=0 and Dec=0 were entered.
2. Enter a search radius (about the specified coordinates) in Box 1. Also be sure to select the desired units, which are most likely to be *degrees* for DC3b data.
3. Click the **Find Datasets** button; the query may take several seconds, depending on the search radius. A selection of available image catalogs will appear (on the left, below the dashed line in Figure 2-1), which include ImSim data with separate entries for raw and calibrated images.
4. In this case, the LSST ImSim calibrated science images were queried by clicking the highlighted text. This action populates the lower right-hand portion of the panel, which is a table of image metadata, one row for each image that matches the search criteria.
5. The right-most column of the table contains a hyperlink to the image; clicking on that text will download the file to your local disk. The filename will be as described in Section 2.1.2, except that the directory delimiters will be replaced with underscores.

Method 2: Exposure Catalog and wget

An alternative method for downloading images involves creating a list of desired files, and using the **wget**[9] software to download them. This software will (by default) preserve the directory structure of the originating file system. The list can be created by accessing the Science Exposure catalog through the **Gator** interface (see section 2.2.1.), constructing a query that selects the image of interest, and downloading the list of URLs to your local machine. The process is described in detail below.

1. In the **Gator** interface (see Figure 2-2), click the `Science CCD Exposure Metadata` catalog from the top-level page.
2. Use the query field limits to enter qualifiers as necessary. In this example, we wish to query for images observed in *r*-band, that are within 30 arcmin of RA=0, Dec=0. Then click the **Run Query** button. It is best to do a practice run of the query and display appropriate fields (RA, Dec, filter name, etc.) to verify that it returns the results that were intended.
3. For the last iteration, uncheck all of the boxes in the **Sel** field except **url** and re-run the query, which will generate a catalog that contains only the URLs of the desired images. Either download the table (if it is large), or view the table directly and cut/paste the results into an ASCII file (in this example, named `myList.txt`).
4. Use the following command to retrieve this list of files:

```
unix% wget –ri myList.txt -nH --cut-dirs=2
```

[9] The **wget** software is available at: http://www.gnu.org/software/wget/. The user manual is available at http://www.gnu.org/software/wget/manual/wget.html.

Using **wget** will also preserve the directory structure through the path (unless explicitly disabled with a command-line argument). Users who wish to download large numbers of images (many focal planes, say) should consult with DM staff for alternative methods of mass-replicating data.

2.3. Software Resources

The astronomical community is fortunate to have access to a wide variety of software applications, tools, services, and software languages with which to discover, access, and analyze data. There is nothing so special about LSST data products that would preclude the use of most software that astronomers and engineers routinely use for analysis. However, the DM team has experience with some packages that have proved especially useful, which we summarize in Table 2-7 below. All of the listed software is free. We also include some sub-packages or software libraries under the parent package, if applicable, and if they would need to be installed separately on your computer. This list is by no means exhaustive, and it is mainly focused on data retrieval and analysis, with less emphasis on supporting software development by users. You may find additional suggestions for useful software on the User Forum or on the Science Wiki.

Table 2-7: Applicable Software Packages

Package/ Sub-Package	Version	Description
Aladdin	7.015b	Image display and analysis tool. Available at: http://aladin.u-strasbg.fr/
SAOimage/DS9	6.2	Image display and analysis tool. Available at: http://hea-www.harvard.edu/RD/ds9/
pyds9	1.1	Python interface to XPA to communicate with DS9 .
IRAF	2.14+	Widely used astronomical image analysis software. Available at: http://iraf.net/
TABLES	3.12	IRAF package for construction and analysis of tabular data. Available at: http://www.stsci.edu/resources/software_hardware/tables
python	2.6.6	Programming language useful for general scientific analysis. Also used in LSST DM software stack. Available at: http://python.org/
atpy	0.9.4	Python package for manipulating tabular astronomical data in a variety of formats, including FITS, VOTable, ASCII, as well as accessing SQLite, MySQL, and PostgreSQL databases. Available at: http://atpy.sourceforge.net/
matplotlib	1.0.1	Currently the most generally capable plotting library for python. Available at: http://matplotlib.sourceforge.net/
NumPy	1.6.0	Numerical operations using arrays. Available at http://sourceforge.net/projects/numpy/files/
PyFITS	2.4.0	Python package for creating and updating FITS images and tables. Available at: http://www.stsci.edu/resources/software_hardware/pyfits
TopCat	3.8	Tabular data display, editor, and analysis application. Available at: http://www.star.bris.ac.uk/~mbt/topcat/
wget	1.12	GNU package for retrieving files using HTTP, HTTPS, and FTP. Available at: http://www.gnu.org/software/wget/

More ambitious users who wish to work with the LSST processing software will find the developer guide by Wittman, et al. (2010) extremely helpful.

2.4. References and Further Information

Contributing Authors

Dick Shaw, Ray Plante, and K-T Lim contributed to the content of this chapter.

References

Pence, W. D., Chiappetti, L., Page, C. G., Shaw, R. A., & Stobie, E. 2010, *Definition of the Flexible Image Transport System (FITS), Version 3.0*, A&A, 524, A42; also available at: http://fits.gsfc.nasa.gov/fits_standard.html

Pence, W. D., Seaman, R., & White, R. L. 2009, *Lossless Astronomical Image Compression and the Effects of Noise*, PASP, 121, 414

Szalay, A. S., Connolly, A. J., & Szokoly, G. P. 1999, AJ, 117, 68

Wittman, D., Abbott, D., Bickerton, S., et al. 2010, *Scientists' Guide to the LSST Applications Software* (Tucson, AZ: LSST Corp.), available at: http://dev.lsstcorp.org/trac/attachment/wiki/Applications/lsstManual.pdf

Chapter 3: Input Data

The data that were processed by the LSST data processing software for the current data challenge have two sources: high fidelity simulations of the LSST telescope and camera, and data from the CFHT Legacy Survey. This chapter provides some background on how these data were produced, and the telescopes and instruments that were either simulated or operated to produce them. In both cases the raw data were organized (or re-organized) prior to processing in a way similar to what will be presented to the Data Management processing system—i.e, partitioned at the "amplifier" level, in order to leverage the highly data-parallel nature of the pipelines.

3.1. Image Simulation Data

The LSST image simulation software provides high fidelity, end-to-end simulations of the sky. These simulated images and catalogs extend to r =28 (deeper than the expected 10 year depth of the LSST co-added images) and are intended to be used for: designing and testing algorithms for use by the data management groups, evaluating the capabilities and scalability of the reduction and analysis pipelines, testing and optimizing the scientific returns of the LSST survey, and providing realistic LSST data to the science collaborations to evaluate the expected performance of LSST. Connolly et al. (2010) describe the simulations and the supporting software in detail. Here we summarize the main features of the simulations.

3.1.1. Features of the Simulations

The simulation of LSST images is divided into three primary components: a database of simulated astronomical catalogs, software for generating an instance catalog of sources based on a pointing at a particular epoch, and software for simulating LSST images based on the input catalog, atmospheric conditions and the telescope/camera system.

Base Catalog

The fundamental database of astronomical objects, called the *base catalog*, is derived from a variety of models of different astronomical phenomena. These include N-body cosmological simulations, models for Galactic structure, simulations of solar system objects, characterizations of transient and variable phenomena, extended sources (galaxies), and an interstellar extinction model. For all categories of sources, the database includes characterizations of their attributes: spectral, photometric (brightness and variability), astrometric (proper motions and distribution on the sky), and morphological properties (including light distribution for extended sources).

Several types of variability are included in the catalog, and are listed in Table 3-1 below along with the type key that is recorded in the ImSim reference database.

Table 3-1: Types of Variable Sources in ImSim

ID Code	Type of Variable
1	RR Lyrae
2	Active galactic nucleus
3	Lensed quasar
4	M-dwarf flare

ID Code	Type of Variable
5	Eclipsing binary
6	Short-duration microlensing
7	Long-duration microlensing
8	AM CVn
9	Cepheid

For the galaxy models the redshift and magnitude distributions approximate those observed by deep imaging and spectroscopic surveys. The internal extinction for the galaxies is inclination dependent and a fraction of the galaxies contain a central AGN. The AGN have a variability model that is wavelength dependent and matches the parameters observed for the SDSS stripe 82 imaging data. The host galaxy is selected to have the closest match to the preferred stellar mass and color at the AGN's redshift. Each galaxy hosts at most one AGN. In the final catalog, AGN host galaxies have high stellar masses (~10^9 to 10^{11} M_O) and cover a range of colors, in general agreement with recent studies of the host galaxy population (e.g., Xue et al. 2010). Within the galaxy sample are a series of strong lenses (with appropriate time delays). The stellar populations have been extended to include RGB, BHB and RRLy stars. The sampling of these stars in color space has been improved with a new interpolation method and the color match to M-, L-, and T-dwarf colors is a better fit to SDSS observations. There is a known discrepancy between the *u-g* colors in the simulations and those observed for the SDSS.

Instance Catalog

The base catalog is queried using sequences of observations that are derived from the ***Operations Simulator***, which models the sky coverage based on models of observing cadence, weather, lunar cycles and seasonal effects, telescope performance, etc. These queries generate source positions and brightnesses based upon the modeled spatial distributions. Any given observation includes specifications for the telescope pointing at a particular time, and observing conditions (seeing, sky brightness, lunar phase and angular separation from the observation). For a given pointing, an instance catalog is generated, where source positions are propagated based on parallax and proper motion (for stars), or ephemerides for solar system bodies. Brightnesses are derived using the filter transmission functions after applying corrections for source variability. The resulting source table, or *instance catalog*, is used to generate the simulated image.

Image Simulation

Images are generated by ray-tracing individual photons from the catalog of sources through the atmosphere, and the telescope and camera optics. The distribution of photon energies is drawn from the source spectral energy distributions (SEDs) in the instance catalog. The realized image includes a model for the detector physics to convert detected photons to electrons, including bias stability, dark current, read noise, charge traps and transfer efficiency, and a model for the effects of cosmic rays. All of these models are based upon the current set of specifications and tolerances for the telescope, camera, and focal plane detector array. The final, formatted image includes simulation of the focal-plane read-out electronics and detector defects.

Sources and Effects Not Included

Detailed though the image simulations are, there are types of sources and effects that are not yet modeled. Future enhancements to the input catalogs will include incorporating new cosmological simulations that better match observed characteristics of distant galaxies, such as morphology at high

redshift and gravitational lensing; and the inclusion of supernovae, GRBs, and classical novae. The modeling of instrumental effects has yet to include artifacts such as glints, ghosts, electronic cross-talk, and fringing in the near-IR.

3.1.2. ImSim Data Selection

Seven slightly overlapping LSST fields, covering approximately 10 deg^2, each have been simulated with the ImSim software for DC3b. These fields are all contiguous; the field locations are given in Table 3-2. The locations and orientations of the individual exposures for these fields were selected from the fifth year of observations as defined by the Operations Simulator run 3.61. The fields cover roughly 60 deg^2 and fall within a 15°×15° patch centered at RA=0.0, Dec=0.0. The overlapping footprints of the exposures at various orientations on the sky facilitate the removal of gaps in the focal plane, detector artifacts, etc., during the creation of deep stacks.

Table 3-2: LSST Standard Simulated Fields[10]

Field	Field Center				Extent	N_{Images}
	RA	Dec	l^{II}	b^{II}		
2426	0:06:14	–04:48:34	95.2	–65.2	3.°5	64 (g:10 r:18 i:20 z:11 y:5)
2536	0:00:00	–02:18:29	94.4	–62.3	3.°5	69 (g:7 r:20 i:26 z:12 y:4)
2544	0:12:34	–02:15:53	100.8	–63.5	3.°5	56 (g:8 r:11 i:23 z:9 y:5)
2656	0:06:20	+00:11:15	99.4	–60.6	3.°5	64 (g:9 r:18 i:23 z:10 y:4)
2762	0:00:00	+02:38:06	98.3	–57.8	3.°5	65 (g:5 r:24 i:22 z:9 y:5)
2770	0:12:45	+02:41:17	104.1	–58.8	3.°5	67 (g:11 r:23 i:24 z:8 y:1)
2886	0:06:26	+05:05:25	102.6	–56.0	3.°5	65 (g:9 r:20 i:23 z:11 y:2)

Note that roughly half of the regular ImSim images were intentionally created with grey (cloud) extinction turned off. Any data quality assessment for which the effects of weather would be an unwelcome complication can instead make use of cloudless images.

3.1.3. LSST Camera

The layout of the planned LSST focal plane array (***FPA***) is shown in Figure 3-1. The 189 science CCD detectors will be mounted onto a total of 21 ***raft*** physical structures containing 3×3 detectors each. The rafts are mounted in the focal plane in a regular grid of 5×5, but with the corners replaced by separate sensors for telescope guiding and wavefront sensing.

[10] The list of simulated fields will grow substantially over the course of the current data challenge.

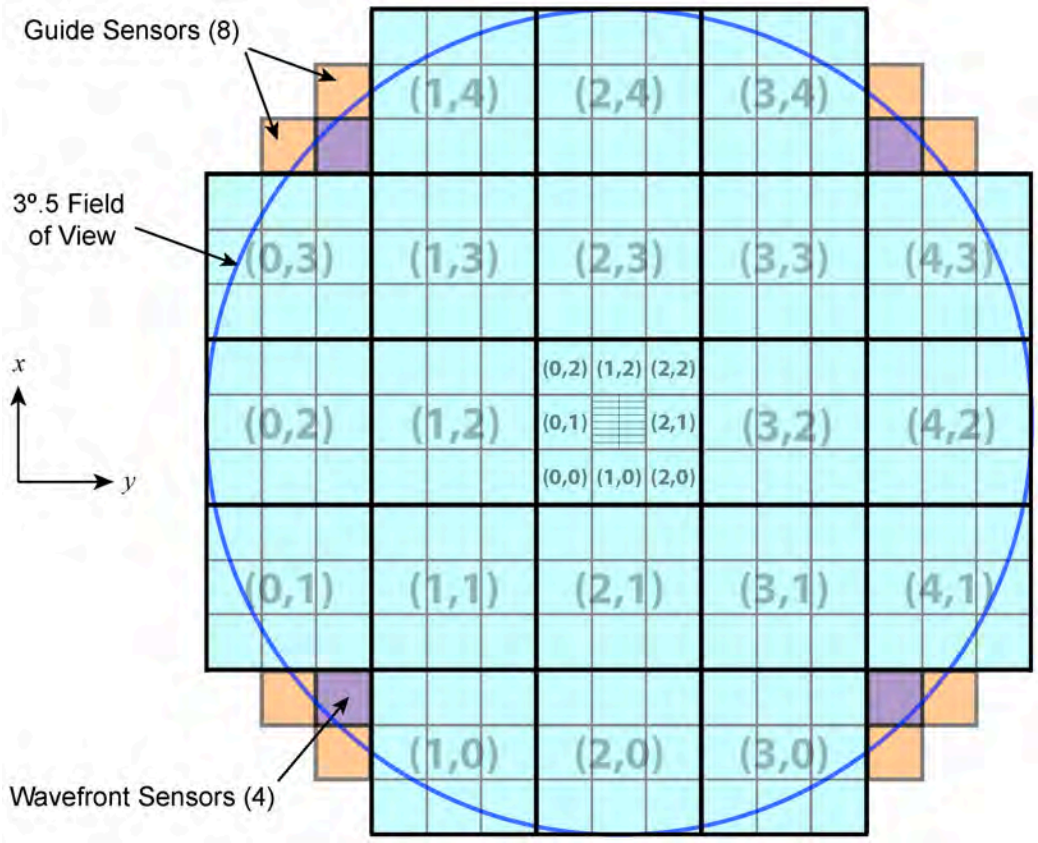

Figure 3-1 Geometry of the planned LSST focal plane array. The rafts (*blue squares*) are identified by pairs of integers to indicate their location in (*x, y*) within the FPA, with (0,0) denoting the lower-left corner. Individual sensors are similarly identified with respect to their location on a particular raft, as shown for the central raft in this figure; individual amplifiers are shown for the central sensor. Also shown are the locations of sensors on the periphery of the FPA that will be used for guiding and wavefront sensing. The extent of the 3°.5 diameter field of view is indicated (*blue circle*).

The characteristics of the sensors in the simulated FPA are summarized in Table 3-3. These values are taken from the design specifications, and may not reflect the as-delivered camera in every respect.

Table 3-3: Characteristics of Simulated LSST CCDs

Sensor Dimensions	Photo-active area: 4072 × 4000 pixel (13′.3 × 13′.6)
Amplifiers	16 (arranged 2 × 8)
Pixel size	10 μm (0.20 arcsec @ field center)
Gain	~1. 7 e^-/ADU
Read Noise	~3.4 e^-
Dark current	2 e^-/pixel/s
Saturation	~57,000 ADU (~96,600 e^-)
Full well	100,000 e^-
FPA gaps	
between CCDs	150 pix (30") in *x*, 215 pix (43") in *y*
between Rafts	235 pix (47") in *x*, 185 pix (37") in *y*

The geometry of a single LSST sensor is illustrated in Figure 3-2. Note that the LSST telescope mount is an Alt-Az design, so that the world coordinates (i.e., RA and Dec) will have an orientation on the FPA that in general varies from visit to visit, depending upon the pointing of the telescope and the rotation angle of the camera.

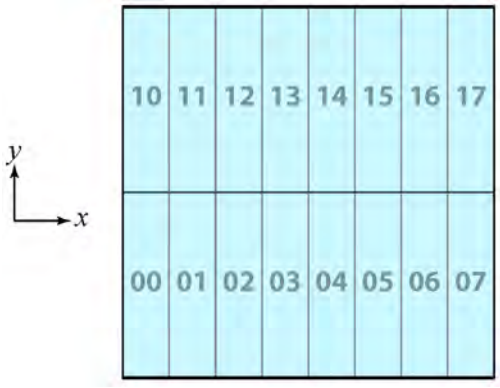

Figure 3-2: Geometry of a single CCD in the LSST focal plane, with amplifier designations and sensor coordinate system indicated. For ImSim data, each 513×2001 pixel amplifier in a raw image includes one row and 4 columns of virtual overscan.

The (ideal) passbands that are planned for LSST are shown in Figure 3-3. They approximate those that were used for the Sloan Digital Sky Survey (SDSS), with the addition of a *y*-band (designated y_4 here to differentiate it among the four alternative curves currently under consideration) in the near-infrared.

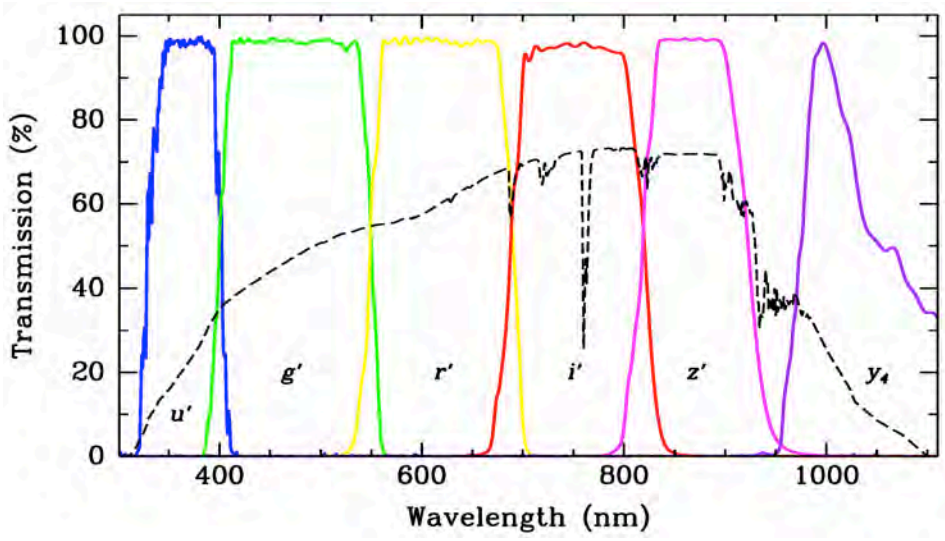

Figure 3-3: Transmission of the filters planned for LSST (*colored curves*), along with the anticipated response of the telescope + camera system multiplied by unit atmospheric transmission (*black curve*).

3.2. CFHT Legacy Survey Data

The second collection of input data that were used in the DC3b processing originate from the CFHT Legacy Survey (CFHT-LS), which was a major photometric survey of several fields that was conducted from 2003—2008. The CFHT-LS is in many ways similar to the planned LSST survey, including many epochs over a substantial span of time, a nearly identical filter set, similar spatial resolution and scale, plus all the foibles of a real observing program. The CFHT-LS survey was

described by Cabanac, et al. (2007), and a great deal of documentation is available on the project website[11].

3.2.1. CFHT Data Selection

All raw data from the CFHT-LS *Deep* and *Wide Synoptic* fields will be processed for DC3b. This includes all survey images obtained with MegaCam between 2003 May 23 and 2008 Feb 16. The fields are all far from the Galactic plane, as described in Table 3-4; the column labeled "notes" provides information about data from other, spatially overlapping surveys. The Deep fields are approximately the same size as the MegaCam field of view (FoV). Individual exposures for these fields were spatially dithered by small offsets to facilitate the removal of gaps in the focal plane, detector artifacts, etc., during the creation of the final, stacked frames. The footprints of the Wide fields required both dithering and tiling the individual exposures.

Table 3-4: CFHT-LS Survey Fields

Field	Field Center				Extent	N_{Img}	Notes
	RA	Dec	l^{II}	b^{II}			
Deep 1	02:26:00	−04:30:00	172.0	−58.0	1° × 1°	3151	Lies within W1
Deep 2	10:00:29	+02:12:21	236.7	+42.0	1° × 1°	2746	On the COSMOS/ACS survey field
Deep 3	14:17:54	+52:30:31	96.3	+59.7	1° × 1°	3528	Lies within W3
Deep 4	22:15:31	−17:44:05	39.2	−52.8	1° × 1°	2998	Around the quasar LBQS2212-17
Wide 1	02:18:00	−07:00:00	172.5	−61.2	8° × 9°	2855	On the XMM LSS field
Wide 2	08:54:00	−04:15:00	232.1	+24.7	7° × 7°	1145	
Wide 3	14:17:54	+54:30:31	98.8	+58.4	7° × 7°	1956	On the Groth Strip
Wide 4	22:13:18	+01:19:00	63.2	−42.5	4° × 4°	974	On the VVDS 22h & UKIDSS DSX fields

Partly by design, and by the desires of follow-on survey teams to maximize their science impact, the CFHT-LS survey fields overlap a number of major survey areas. These surveys span a large range in wavelength and spatial resolution, such as the AEGIS survey of the Groth Strip (Davis, et al. 2007), the COSMOS survey (Scoville, et al. 2007), the XMM-LSS Survey (Pierre, et al. 2009), and the VIMOS VLT Deep Survey (Le Fèvre, et al. 2005). Together with the T0005 Terapix deep stacks and object catalogs for CFHT-LS (Mellier, et al. 2008), these surveys greatly extend the scientific potential of the CFHT-LS, but also serve as important points of comparison for evaluating the quality of the LSST production processing.

CFHT-LS data have not been processed at productions scale and are not currently available. These data products are planned for a later phase of DC3b.

3.2.2. MegaCam

The CFHT-LS was carried out with the wide field optical imaging camera *MegaCam*, which features a focal plane array with 36 CCDs that cover a roughly 1°×1° FoV. The geometry of the MegaCam

[11] See http://www.cfht.hawaii.edu/Science/CFHLS/

focal plane array is shown in Figure 3-4; note that since CFHT is an equatorial-mount telescope, the orientation of the camera on the sky is fixed. The FPA geometry is closely tied to the organization of the image data in the raw FITS files as they are stored in the CADC archive[12]. Specifically the data are stored in FITS Multi-Extension format (see Chapter 1.1), with one extension per CCD. The detectors are each read out using two amplifiers in parallel, denoted *A* or *B* in the figure. The raw images also include the overscan pixels.

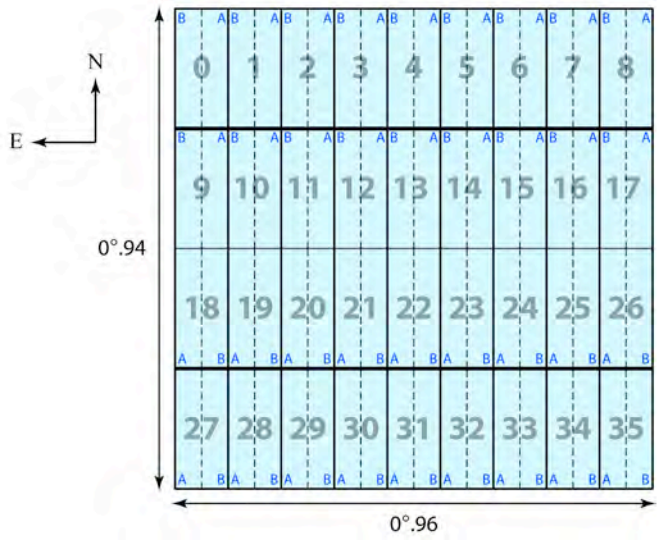

Figure 3-4: Geometry of the focal plane array for MegaCam. The size of the FoV is shown, and sky orientation is indicated (*upper left*). The index of the 36 CCDs runs sequentially from upper left to lower right; amplifiers are labeled *A* or *B* near the read-out origin of each sensor.

The detectors have very good cosmetic and performance characteristics; a summary is provided in Table 3-5.

Table 3-5: Typical Characteristics of MegaCam CCDs

Array Dimensions	Photo-active area: 2048 × 4612 pix (6′.4 × 14′.4)
Amplifiers	2
Pixel size	13.5 μm (0.187 arcsec @ field center)
Gain	~1.67 e^-/ADU
Read Noise	~5 e^-
Dark Current	<2 × 10^{-3} e^-/s/pix
Linearity	within 0.1% below saturation
Saturation	65536 ADU (~110,000 e^-)
Full well	~150,000 e^-
CCD gaps	
Small gaps	70 pix (13 arcsec)
Large gaps	425 pix (80 arcsec)

The passbands that were used for CFHT-LS are shown in Figure 3-5. They approximate those that were used for the SDSS, except that the *u*-band (designated *u**) is slightly broader.

[12] See the CFHT MegaCam raw data description at http://www.cfht.hawaii.edu/Instruments/Imaging/MegaPrime/rawdata.html for details.

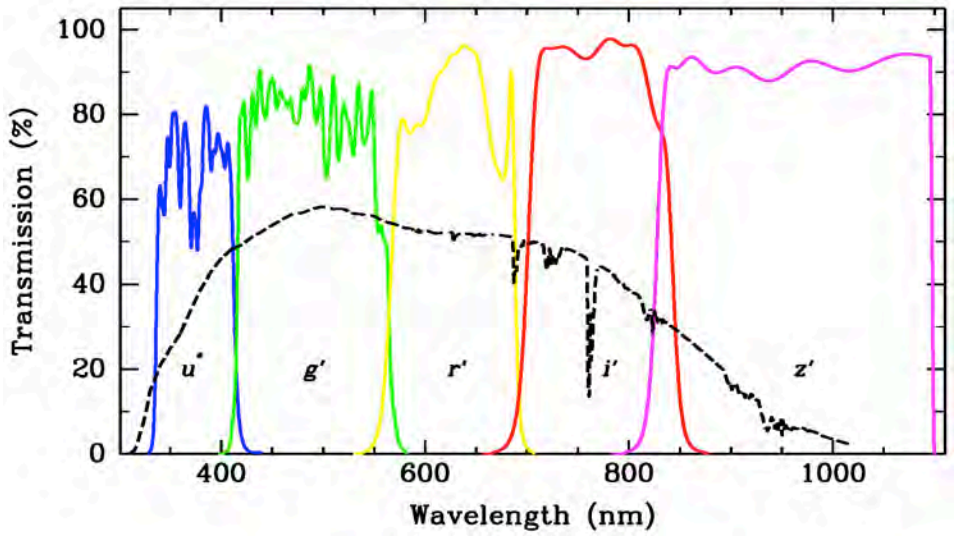

Figure 3-5: Transmission of the filters used for the CFHT-LS (*colored curves*), along with the response of the telescope + camera, multiplied by unit atmospheric transmission (*black dashed curve*).

3.3. References and Further Information

Contributing Authors

Most of the information about the CFHT-LS data was taken or derived from the project web site, and from the T0005 Release Document (Mellier, et al. 2008). Andrew Connolly, Simon Krughoff, and John Peterson provided the information on the Image Simulations.

References

Cabanac, R. A., et al. 2007, *The CFHTLS Strong Lensing Legacy Survey*, A&A, 461, 813

Connolly, A. J., et al. 2010, *Simulating the LSST System*, Proc. SPIE, 7738, 53

Davis, M. et al. 2007, *The All-Wavelength Extended Groth Strip International Survey (AEGIS) Data Sets*, ApJ, 660, L1

Le Fèvre, O., et al. 2005, *The VIMOS VLT Deep Survey. First Epoch VVDS-Deep Survey: 11,564 Spectra with 17.5≤IAB≤24, and the Redshift Distribution Over 0 ≤ z ≤ 5*, A&A, 439, 845

Mellier, Y., et al. 2008, *The CFHTLS T005 Release*, (Paris: Institut d'Astrophysique de Paris). Available at: http://terapix.iap.fr/cplt/oldSite/Descart/CFHTLS-T0005-Release.pdf

Pierre, M. et al. 2004, *The XMM-LSS Survey. Survey Design and First Results*, JCAP, 9, 11

Regnault, N., et al. 2009, *Photometric Calibration of the Supernova Legacy Survey Fields*, A&A, 506, 999

Scoville, N., et al. 2007, *The Cosmic Evolution Survey (COSMOS): Overview*, ApJS, 172, 1

Xue, Y. Q., et al. 2010, *Color-Magnitude Relations of Active and Non-active Galaxies in the Chandra Deep Fields: High-redshift Constraints and Stellar-mass Selection Effects*, ApJ, 720, 368

For Further Reading

Additional details of the LSST system design, expected performance, the planned operations model, and the expectations for scientific discovery may be found in the LSST Science Book, which is available at http://adsabs.harvard.edu/abs/2009arXiv0912.0201L. The science requirements for the LSST hardware and survey are described by formal LSST documents; a detailed discussion of these requirements and their realization in system requirements is presented in:

Ivezic, Z., et al. 2008, *LSST: From Science Drivers to Reference Design and Anticipated Data Products* (astro-ph/0805.2366), available at: http://lsst.org/files/docs/overview_v1.0.pdf

Chapter 4: Data Processing and Calibration

Science data processing for LSST is organized into a series of ***productions***, which are episodes of processing that are organized to achieve a particular purpose, such as issuing event alerts during a night's observing; generating calibrated images and object catalogs on an annual basis; or constructing calibration reference products. The types of productions that have been identified so far are given in Table 4.1, and in general they operate over a particular timescale of relevance. The focus of the current data challenge (DC3b), and of this *Handbook*, is the *data release production* (i.e., the annual effort). Each production is a fairly involved activity that takes multiple data inputs and executes software on a massively parallel computing platform in order to generate one or more science data products, related metadata, and data quality information. The software is organized into a series of ***pipelines***, or independently executable codes, each of which consists of one or more logical ***stages*** that perform discrete algorithmic operations.

Table 4-1: Types of Data Productions

Production	Timescale	Description
Alert	Nightly	Performs basic calibration and difference image analysis within 60s of shutter close to detect objects that are unknown or that have changed in brightness by a predefined amount, relative to a template image of the sky. Basic quality screening is performed prior to issuing an ***alert*** to the community, which will include coordinates, a measured brightness, as well as image cut-outs of the target and template images.
Moving Object	Daily	Associates sources with either known moving objects, or sets of ***tracklets*** for unknown sources, in an attempt to identify solar system objects. Constructs or refines orbits for moving objects and stores the information in the science database.
Calibration	~Monthly	Combines closed-dome calibration images (bias, dark, flat-field) into calibration reference images, calibrates relative system throughput; calibrates on-sky measurements of amplifier cross-talk, grey extinction, scattered light, etc.; combines time-dependent positional information from source catalog to calibrate astrometric reference system; computes global photometric calibration; etc.
Data Release	Annually	Complete reprocessing of all images obtained up to a particular cut-off date. Generates calibrated visit images, image templates, deep detection stacks, and object, source, and moving object catalogs.

The computations are optimized for high throughput on highly parallel computing platforms, but algorithmic software is organized to be largely independent of the execution environment. This chapter will focus on *what* scientific operations are performed in the Data Release Production for DC3b, with relatively little discussion about *how* they are performed except to the extent that they affect the organization of the output data.

4.1. Pipeline Processing

4.1.1. Overview

The flow of the science data through the initial stages of the pipeline processing is shown in Figure 4-1 for DC3b. Each step of the processing, indicated by the boxes in the center of the figure, is

described in detail in the following subsections. As explained below, some steps are not performed for data where the correction in question is either not needed, or the functionality has not yet been implemented in the pipeline. Inputs to the processing include the raw science frames, configuration files, calibration reference images, and catalogs. Outputs include the various reduced science images, including their concomitant data, catalogs, and data quality metadata. Intermediate products that are produced during the course of pipeline processing, but that are not archived, are not shown.

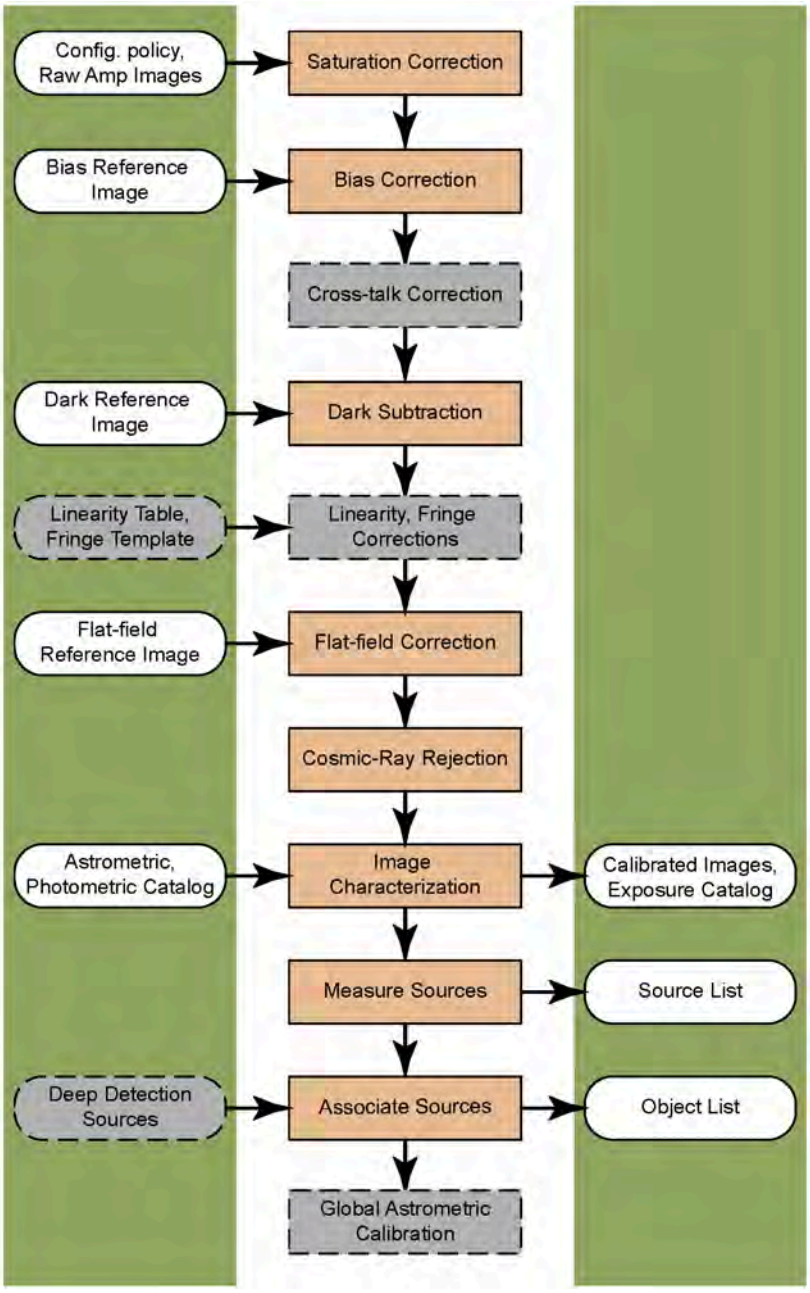

Figure 4-1: Flow of data taken with a common filter through single-visit processing of the Data Release Production pipeline to produce Level 2 data products. Images and catalogs are shown as inputs to (*left column*) or outputs of (*right column*) the processing. Boxes in grey are placeholders for functionality that has not yet been implemented. See Figure 4-2 for down-stream processing stages.

4.1.2. Amplifier-level Processing

The following steps are performed in parallel on amplifier-level image subsections.

Saturation Correction

At the start of pipeline processing the pixel values are examined to detect saturation (which will naturally also identify bleed trails near saturated targets, and the strongest cosmic rays). These values, along with pixels that are identified in the list of static bad pixels, are flagged in the data quality mask of the science image. (The list of all pathologies that are tracked in each mask plane is given in Table 2-5.) All pixels in the science array identified as "bad" in this sense are interpolated over, in order to avoid problems with source detection and with code optimization for other downstream pipeline processing.

Interpolation is performed with a linear predictive code, as was done for the Sloan Digital Sky Survey (SDSS). The PSF is taken to be a Gaussian with sigma width equal to one pixel when deriving the coefficients. For interpolating over most defects the interpolation is only done in the *x*-direction, extending 2 pixels on each side of the defect. This is done both for simplicity and to ameliorate the way that saturation trails interact with bad columns.

Bias Correction

The bias correction is applied to remove the (additive) electronic bias that is present in the signal chain. The bias is to first approximation a constant pedestal, but it has low-amplitude structure that is related to the electronic stability of the bias during read-out of the detector segment. The processing pipeline removes the bias contribution in a two-step process. In the first step, the median value of non-flagged pixels in the over-scan region is subtracted from the image. In the second step, the reference bias image is subtracted from the science image to remove higher-order structure. Following the bias correction, the pixels are scaled by the gain factor for the appropriate CCD. The brightness units are electrons (or equivalently for unit gain, detected photons) for calibrated images.

Cross-Talk Correction

No detectable amount of cross-talk occurs between the various readout channels of the CCDs in the CFHT-LS data. It is not known whether cross-talk will be a factor in the LSST camera, and this effect is not modeled in the ImSim images. The effect, when present, is to introduce a small fraction of the signal from one CCD into the signal chain of the CCD that shares the same electronics, such that "ghosts" of bright objects appear in the paired CCD. This is an additive effect, and is most noticeable for sources that are at or near saturation. The pipeline has a placeholder for this correction, should it be necessary, **but no cross-talk correction is implemented at this time**.

Dark Correction

The dark current, i.e., the signal introduced by thermal electrons in the detectors with the camera shutter closed, is significant for the ImSim images, which reflects a rather loose specification on the LSST detector performance. Dark correction is applied by subtracting a reference Dark calibration frame that has been scaled to the exposure time of the visit image. On the other hand, the **dark current is extremely low for the CFHT-LS data, so no dark correction is applied** in that case.

Linearity Correction

The response of the CCD detectors to radiation is highly linear for pixels that are not near saturation, to better than 0.1% in the case of CFHT-LS data; non-linearity is not enabled in the ImSim data. **Currently, no linearity correction is applied in the pipelines.** Were a correction necessary it would likely be implemented with a look-up table, and executed following the dark correction but prior to fringe correction.

Fringe Pattern Correction

A fringe pattern is evident in CFHT-LS data that were obtained with the reddest filters: the *i′*-, *z′*-, and *y*-bands. The pattern occurs because of interference between the incident, nearly monochromatic light from night sky emission lines (both from air glow and reflected city lights) and the layers of the CCD substrate. The details of the fringe pattern depend mostly upon the spatial variation in thickness of the top layer of the substrate, but also depend upon a number of other factors including the wavelength(s) of the incident emission lines, the composition of the substrate, the temperature of the CCD, and the focal ratio of the incident beam. The amplitude of the fringe pattern background varies with time and telescope pointing.

No fringe pattern correction is currently implemented in the DM processing pipeline. In the case of ImSim data, the omission is benign because fringing is not simulated. In the case of CFHT-LS data, the amplitude of the fringes can be large compared to the mean sky background, and is ~6% in *i′* and ~15% in *z′*.

Flat-Field Correction

The flat-field correction removes the variations in the pixel-to-pixel response of the detectors. The flat-field is derived for each filter in one of two ways, depending upon the data source: for CFHT-LS the flat-fields are generated from images of the twilight sky; for ImSim the flat-fields are generated from a simulated continuum source. In both cases the flat-field corrects approximately for vignetting across the CCD. The flat-field correction is performed by dividing each science frame by a normalized, reference flat-field image for the corresponding filter.

4.1.3. CCD-level Processing

The following steps are performed after the contiguous amplifier subsections are assembled into a CCD-level image.

Background Subtraction/Cosmic Ray Rejection

The background in science images is removed prior to cosmic ray rejection. The background level has multiple origins, including twilight, airglow, scattered light (from the moon), ghosts and glints from the optical surfaces of the telescope and instrument optics, and the extended wings of very bright stars. The low-spatial frequency scattered light is determined in multiple, non-overlapping regions across each CCD. Within each region a statistic is computed to estimate the local background, and the resulting values are interpolated over the extent of the CCD. The statistic, the interpolation scheme, and the region sizes are all configurable, and are currently: clipped mean, a natural spline, and 256 × 256 pixels, respectively.

Cosmic rays (CRs) are detected on single exposures as spatially compact, relatively bright sources that are not similar in shape to the point-spread function (PSF). The current algorithm is very similar to that used in the SDSS pipeline, which searches pairs of pixels for gradients in the image that are too steep to have been generated by a PSF. After a first round of identifying pixels associated with these steep gradients they are assembled into individual CRs, which are required to have a minimum (configurable) number of counts. The pixels around the CRs are then examined (with slightly relaxed criteria) and potentially added to the CR; this process is iterated. Finally, a flag is set in the quality mask of the calibrated images for all pixels that are identified with CRs, and the corresponding pixels

in the individual science frames are then interpolated over. The visit exposures are then averaged to create a single science image.

The combination of pairs of images in a visit is disabled for the present, pending a resolution of large (approaching 1 arcsec) apparent motions of the field between the component ImSim exposures. This limitation also disables the intended strategy of using image pairs for an additional CR detection pass.

Image Characterization

The background-subtracted visit images are now ready for single-frame calibrations and basic measurements. Note that all calibrations and measurements at this stage of the processing are performed at the CCD level. Therefore measurements of sources that fall on or near detector boundaries will be affected.

World Coordinate System Calibration The WCS calibration for science images is described by a two-dimensional polynomial (the function type and coefficients are found in the header) of a tangent-plane projection of stellar coordinates to the image pixel grid. The lower-order terms relate to the location of the reference pixel on the sky, the plate scale, and the rotation of the image. Higher-order terms, up to fourth-order, may be added to model the distortion from the optical system and differential atmospheric refraction. These distortion terms follow the Simple Image Polynomial (SIP) convention for representing distortions in FITS format (Shupe & Hook 2008). The magnitude of the terms in the distortion function varies with the filter and with the airmass of the observation. **Note: at present there is no constraint on the WCS coefficients (e.g., position, rotation, plate scale, or size and order of the SIP distortions) from one CCD to another across the focal plane.**

The per-CCD WCS solution is based on the Astrometry.net code (Lang, et al. 2010). Briefly, the approach is to find asterisms composed of a few stars (typically 4) in the image, and search for similar asterisms (i.e., with similar relative geometry, invariant to position, rotation, and scale) in a reference catalog. This generates hypotheses about where the image might be on the sky. Each hypothesis is checked by predicting where other stars should be found, and evaluating this prediction using Bayesian decision theory. If the image already has a complete WCS, it is possible (though not yet implemented in the pipeline) to skip the first stage and go straight to evaluating whether it is correct. Knowledge about the plate scale and an estimate of the pointing, which is gleaned from the observing environment, can be used to constrain (and thus speed up) the search. In any case, since the pattern of galaxies in the ImSim input object catalog repeats every ~4°.5, the telescope pointing gleaned from the raw image header is used to rule out solutions that lie outside the FoV. After the low-order terms are solved, a least-squares method is applied to test whether higher-order SIP terms improve the solution.

PSF Estimation The size and shape of the point-spread function (PSF) is determined from well isolated, relatively bright stellar sources across the ***visit*** image focal plane. The PSF shapes are used in down-stream processing, including PSF photometry of all sources and discrimination of stars from extended objects. The characterization of the PSF includes the following steps:

1. From a list of candidate detected sources, measure their properties and retain only those with fluxes that exceed a relatively bright (configurable) threshold.

2. Measure the shapes of these objects using ***adaptive Gaussian moments*** (see below), which is equivalent to fitting a 2-D Gaussian to the brightness profiles of all sources and adopting these moments to represent the source shapes.
3. Find the mode of the distribution of second moments (M_{xx}, M_{yy}) of the sources and exclude those that deviate significantly from the central locus (which are assumed to be populated by point sources).
4. Choose the brightest PSF candidates over a regular spatial grid on the CCD, and perform a principle-component analysis (PCA) decomposition, retaining a small number of eigen images.
5. Use the spatial model of the PSF variation derived above to reject deviant PSF candidates.
6. Repeat steps 3—5 until the rejection converges.

In the above procedure, the shape parameters are determined from ***adaptive moments***, or the second moments of the source intensity distribution, measured using a scheme designed to have near-optimal signal-to-noise ratio. From the SDSS documentation[13]: "Moments are measured using a radial weight function that is adapted interactively to the shape (ellipticity) and size of the object. This elliptical weight function has a signal-to-noise advantage over axially symmetric weight functions. In principle there is an optimal (in terms of signal-to-noise) radial shape for the weight function, which is related to the light profile of the source itself. In practice a Gaussian with size matched to that of the object is used, and is nearly optimal. Details can be found in Bernstein & Jarvis (2002)."

These relatively bright stars are also used to determine empirically the aperture correction and its spatial dependence on each image[14]. This step determines the correction needed to measure system magnitudes using a finite aperture. The process is the following:

1. Perform PSF photometry on the sources.
2. Perform aperture photometry on those same stars, using a pre-configured radius (3.0 arcsec).
3. Determine the aperture correction, defined as the ratio of Flux(PSF)/Flux(Aper), using a second-order polynomial to account for the spatial variation across the image.

The aperture correction will be applied to all point sources that are identified in the Source Detection step below.

Photometric Calibration An estimate is made of the magnitude zero-point of each CCD in each visit image by comparing the published magnitudes in a reference photometric catalog to their instrumental magnitudes, applying color transformations as necessary. Currently, the reference photometric catalog that will be used in DC3b processing for CFHT-LS data is USNO-B1.0 (see Monet, et al. 2003). For ImSim data, the catalog is constructed from models of the Galactic distribution of stars and from models of galaxy types, all of which are simulated in the LSST photometric system. Note that the result of the photometric calibration is to populate the science header with keywords, and to populate the exposure table in the science database. The pixel values remain unchanged, and have units of detected photons s^{-1}.

[13] See http://www.sdss.org/dr6/algorithms/adaptive.html for a description of the use of adaptive moments in SDSS.
[14] There is no requirement that the list of bright stars be identical to that used in PSF determination.

The estimation of the photometric zero-points for ImSim images does not attempt to correct for spatially variable grey (i.e., cloud) extinction on scales smaller than a CCD. The variation in amplitude of the modeled grey extinction is of order 2% across the FoV of a single CCD. The effect of omitting these corrections is to inflate the photometric errors by an amount that is small, but not yet accurately characterized.

Source Detection

Following the production of a calibrated, background-subtracted image, it is examined for ***sources***, or astrophysical targets in the field of view. A copy of the image is first convolved with a circularly symmetric Gaussian brightness profile that has the same width as the PSF for that image. Pixels above a configurable threshold in this smoothed image are flagged, and groups of contiguous pixels are measured to determine the centroid locations, fluxes, and shape parameters of (possibly overlapping) sources. Note that this step is limited to detecting targets smaller than ~10 arcmin (i.e., somewhat smaller than the area of sky covered by a single CCD). The shapes are derived from an algorithm that is similar to that used for the SDSS. The resulting measurements (position, brightness, shape, orientation, and errors on those parameters) for the list of all detected sources are recorded in the ***source catalog*** of the science database. Various conditions that may compromise the quality of the source measurements are encoded in the data quality field; their meanings are summarized in Table 4-2 below. The working definition of pathologies that would render a source scientifically useless for downstream analysis is that one or more of the following bits are set: (0x1, 0x200, 0x800)—i.e., the source includes edge pixels, or that the source center is close to interpolated or saturated pixels.

Table 4-2: Meanings of Source Data Quality Flags

Decimal Value	Hex Value	Text Code	Quality Condition Indicated
1	0x**1**	`EDGE`	Source includes pixels within the edge region of a detector—i.e., the half-width of the smoothing filter used for detection, which is typically ~10 pixels.
2	0x**2**	`SHAPE_SHIFT`	While estimating the best-fit Gaussian filter, the derived centroid varied significantly from the initial guess
4	0x**4**	`SHAPE_MAXITER`	The adaptive moments solution required more than the maximum allowed iterations.
8	0x**8**	`SHAPE_UNWEIGHTED`	The adaptive scheme failed to converge, so the moments are unweighted and therefore noisy and unreliable.
16	0x**10**	`SHAPE_UNWEIGHTED_PSF`	The PSF's "adaptive" moments are unweighted. This flag is currently not used.
32	0x**20**	`SHAPE_UNWEIGHTED_BAD`	The source is so noisy that no shape could be determined. The `SHAPE_UNWEIGHTED` flag will also be set.
64	0x**40**	`PEAKCENTER`	Centroid determination failed: derived center is set to peak pixel.
128	0x**80**	`BINNED1`	Source was found in 1×1 binned image. (Larger binning factors will eventually be used to detect extended, low surface brightness sources.)
256	0x**100**	`INTERP`	Source's footprint includes interpolated pixels.
512	0x**200**	`INTERP_CENTER`	Source's centre is close to interpolated pixels.

Decimal Value	Hex Value	Text Code	Quality Condition Indicated
1024	0x**400**	SATUR	Source's footprint includes saturated pixels.
2048	0x**800**	SATUR_CENTER	Source's center is close to saturated pixels.
4096	0x**1000**	DETECT_NEGATIVE	Source was detected as having negative flux (in a difference image), at a significance of at least 5-sigma.
8192	0x**2000**	STAR	Source size and shape is consistent with being point-like.

While saturated sources, and sources found near the edge of the detectors, are included in the source catalog, large sources that fall near and/or span a gap between CCDs may not be properly counted in completeness statistics.

4.1.4. Source Photometry

Once sources have been identified, the flux is measured using multiple techniques, including aperture photometry and adaptive Gaussian moments. For stars (or any angularly compact source with an approximately stellar profile), the photometric calibration step described above assures that aperture and PSF fluxes will agree to good accuracy. For extended sources such as galaxies, the story is more complicated. A robust model fitting code (*Multi-fit*) is being developed for galaxy photometry, which is being designed to fit multiple components at once to complicated sources. Not all of the functionality has been implemented, but for the current release such sources are fit with a linear combination of multiple components: a delta function, an exponential profile, a de Vaucouleur profile, and a second-order shapelet basis. All components are convolved with the PSF model, and the flux is computed as the integral of the model. Note that the radius and ellipticity of the components are not currently fit: these are fixed to a small number of test points by applying a naïve (Gaussian) correction for the PSF to the adaptive moments of the source. However, the inclusion of the shapelet basis effectively allows for small perturbations in ellipticity and radius from the fiducial values.

There is currently no deblending of overlapping sources in the Source and Object catalogs, and no identification of moving objects. The star-galaxy separation with the current software stack is known to be problematic for angularly small galaxies.

4.1.5. Catalog-level Processing

Source Association

Once the source catalog has been generated from all processed images, the source association pipeline identifies the (large) subset that corresponds to multiple detections of individual astrophysical targets. Source association is carried out with the *OPTICS* algorithm (Ankerst, et al. 1999), although the current implementation is equivalent to the *DBScan* algorithm (Ester, et al. 1996). It is one of the most common clustering algorithms used in the science literature[15] and it is very efficient, with a

[15] See the Wikipedia entry for DBSCAN at: http://en.wikipedia.org/wiki/DBSCAN.

runtime complexity of $O(n \cdot \log n)$. The idea is to examine each source in sequence and form ***clusters***, or candidate sources. Clusters can initially be individual sources, but clusters grow when the spatial separation between candidate members and the cluster is small enough. The algorithm is parameterized on the characteristic spatial separation (ε neighborhood) and the minimum number of points (MinPts) required to form a genuine cluster. These parameters are tuned to a given dataset so that the number of false associations is minimized. Preliminary experiments with the sources extracted from ImSim images lead to setting ε=0.5 arcsec, and MinPts=5.

The algorithm operates on the set of all sources falling into a sky-tile. These sources are taken from all the CCDs in all passbands containing at least one raw amp that is within some padding distance P of the sky-tile; P is chosen based on an estimate of the maximum error in the raw WCSes, and is currently ~15 arcsec. The algorithm visits each source S in the sky-tile (in an arbitrary order). If the ε neighborhood of a source S contains at least MinPts other sources, and S has not already been placed into a cluster:

1. Create a new cluster *C*.
2. Add all the ε-neighbors of *S* that do not already belong to a cluster *C*.
3. Recursively perform step 2 for each ε-neighbor *S′* of *S* that has an ε-neighborhood containing at least MinPts other sources.

If the ε neighborhood of S contains less than MinPts other sources, it is called a *noise source* and is discarded. All clusters are stored in the ***object catalog***.

Objects composed of only one source (i.e., one detection in any passband) are in principle allowed, but have been disabled for the present in order to avoid corrupting the object catalog with garbage sources. When the required tuning of this algorithm is better understood, this restriction will be removed.

Note also that the single-visit sources are not associated with external catalogs of astrophysical targets. This functionality is planned for a later data release.

4.2. Calibration Reference Files

Calibration reference images would ordinarily be created by the Calibration Data Products production, which has not yet been developed. To support this data challenge, the reference images were created off-line, with a manual process. For ImSim data, 10 visits each of bias, dark, and flat-field images were created. The visit pairs were pipeline-processed with the appropriate steps, and averaged. The final calibration files were created from the median of the processed images. The images of each type have the following attributes:

- Biases: These zero-second exposures (with rapid read-out) only contain the electronic signature of bias and read-noise; no cosmic rays are simulated.
- Darks: These images were generated with 150 s exposures with no illumination. This increases the incident cosmic rays by a factor of 10 relative to the nominal duration of a science exposure (but they are removed during the stacking process).

- Flat-fields: These images were generated with 15 s exposures of a simulated source with a flat SED, a brightness of 18 mag/arcsec^2, and an illumination pattern that approximates the vignetting of the telescope/camera optics.

For CFHT-LS data, the calibration reference images created by the Elixir[16] pipeline (and archived at CADC) were used after they were re-formatted for input to the Data Release production pipelines.

4.3. Processing Steps Not Yet Implemented

Not all processing pipelines that are planned for inclusion in the Data Release Production have been implemented in DC3b, but most stages have been designed at some level. The processing steps that are necessary to generate stacked image templates and to detect and identify moving objects are shown graphically in Figure 4-2; this functionality is planned for future data releases.

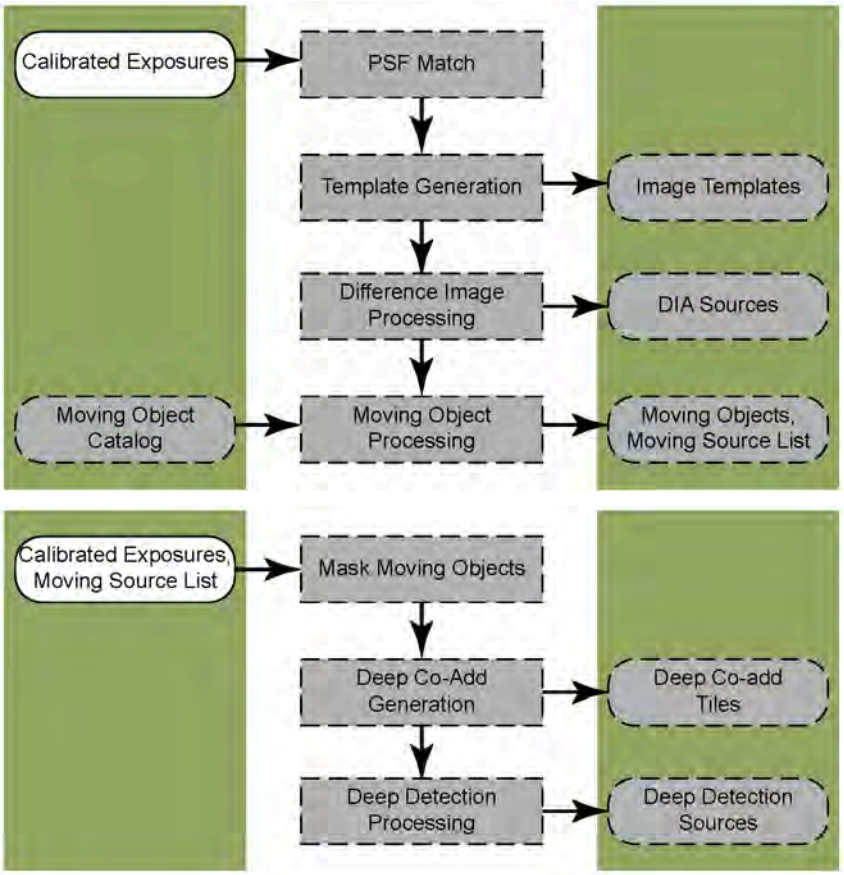

Figure 4-2: Processing flow that follows the steps described in Figure 4-1, which generates Calibrated Exposures for each visit. **These stages have not yet been implemented.** The detection and identification of Moving Objects and the characterization of their orbits (*upper flow diagram*), and the identification of faint sources in deep image stacks (*lower flow diagram*) are planned for a later data release.

Further processing pipelines are being designed that will perform detailed characterizations of partially blended and overlapping sources, identify transient sources, and perform the global

[16] Details of the Elixir pipeline processing for CFHT MegaCam data are described at http://www.cfht.hawaii.edu/Instruments/Elixir/

photometric calibration. The implementation of these pipeline-processing steps is planned for a later data release, and will be described in a revision of this *Handbook*.

4.4. References and Further Information

Contributing Authors

Steve Bickerton, Simon Krughoff, Dustin Lang, K.-T. Lim, Robert Lupton, Serge Monkewitz, John Peterson, Paul Price, and Dick Shaw contributed to the content of this chapter.

References

Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Jörg Sander, J. 1999, *OPTICS: Ordering Points To Identify the Clustering Structure*, in ACM SIGMOD International Conference on Management of Data, ACM Press, 49

Bernstein, G. M., & Jarvis, M. 2002, *Shapes and Shears, Stars and Smears: Optimal Measurements for Weak Lensing*, AJ, 123, 583

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. 1996, *A density-based algorithm for discovering clusters in large spatial databases with noise,* in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, eds. E. Simoudis, J. Han, U. M. Fayyad, AAAI Press, 226

Kantor, J., Axelrod, T., Allsman, R., Freemon, M., & Lim, K.-T. 2010, *Data Challenge 3b Ovewview* (Tucson, AZ: LSST Corp.) Available at https://www.lsstcorp.org/docushare/dsweb/Get/Document-9044/DC3b Scope.pdf

Lang, D., Hogg, D. W.; Mierle, K., Blanton, M., & Roweis, S. 2010, *Astrometry.net: Blind Astrometric Calibration of Arbitrary Astronomical Images*, AJ, 137, 1782

Shupe, D. L., & Hook, R. N. 2008, *The SIP Convention for Representing Distortion in FITS Image Headers*, available at: http://fits.gsfc.nasa.gov/registry/sip.html

For Further Reading

A detailed (and evolving) set of on-line notes on the pipeline stages, and their inputs and outputs, maybe found on the Data Management wiki at http://dev.lsstcorp.org/trac/wiki/DC3bProcessingStages.

Chapter 5: Data Quality Assessment

Data as processed by the DM production software are evaluated for scientific quality by computing various quantities of interest and comparing them, when possible, to established science quality metrics. A detailed set of science requirements for data quality is under development for DM data processing, many of which flow directly from high-level science requirements for LSST as a whole. Shaw et al. (2010) described a DM system approach to LSST science data quality assessment (***SDQA***), which in part consists of defining a variety of metrics against which the data and catalogs will be measured automatically.

Quality is also assessed for ImSim data by comparing measurements produced by the pipelines to the input reference (i.e., "truth") catalog. This chapter describes how to access the automated quality assessment reports, and summarizes the initial assessment of the scientific quality of the generated data products. This chapter will also point out both known problems and specific questions that have yet to be addressed. It is anticipated that members of the Science Collaboration teams will contribute their own analyses to the assessment. This *Handbook* will be updated periodically as the understanding of the released data products matures. It is expected that the analyses themselves will be folded into a definitive DC3b report, to be published separately. This chapter concludes with examples of the types of feedback that the Data Management Team would find most useful.

The automated data quality assessment for DC3b data products presents an excellent opportunity for the Science Collaboration members to explore even deeper questions related to science data quality.

5.1. Assessment of Processed Data

5.1.1. Pipeline Processing Diagnostics

The pipelines (or stages thereof) have the capability to report problems that may occur during processing. Issues not related to algorithmic flaws are generally resolved prior to archiving the data for release. Problematic data (which could result from poor observing conditions; ImSim does simulate such data) are simply flagged. The flags that have been created to date are listed in Table 5-1, a list that will undoubtedly grow when the software recognizes more conditions. These flags reflect quality assessments at the level of a single CCD: see the **`Science_Ccd_Exposure`** table in the database. It may be appropriate to exclude contributions from analysis, depending upon the objective.

Table 5-1: CCD Processing Diagnostics

Decimal Value	Hex Value	Text Code	Quality Condition Indicated
1	0x**1**	`PROCESSING_FAILED`	The pipeline failed to process this CCD.
2	0x**2**	`BAD_PSF_ZEROPOINT`	The PSF flux zero-point appears to be bad.
4	0x**4**	`BAD_PSF_SCATTER`	The PSF flux for stars shows scatter >0.03 mag.

A total of 3102 problem CCDs were identified in the Summer, 2011 data release, and have been flagged in the database (see Table 5-1). Roughly 50% of these CCDs occurred in 13 visits, all of which had spatially variable grey extinction (i.e., clouds) turned on. The cause of the problem is under investigation.

5.1.2. Automated Quality Reports

A quality assessment component of the DM processing system, known as *pipeQA*, is under active development and has already proven useful for validating the processing software. The automatically generated assessments for the current data release are provided through a web interface, which currently generates views for individual visits[17]. The results, or data quality artifacts, consist of summary statistics, plots, and reports of specific tests against quality thresholds. The information is substantial, and will only be briefly summarized here; team members are encouraged to explore the results in detail. The artifacts for the current data release are available to Science Collaboration members at http://lsst1.ncsa.uiuc.edu/pipeQA/public/. The initial screen (home page) presents summary statistics for the processed images, followed by a list of visits that have been processed for the Data Release, similar to that in Figure 5-1.

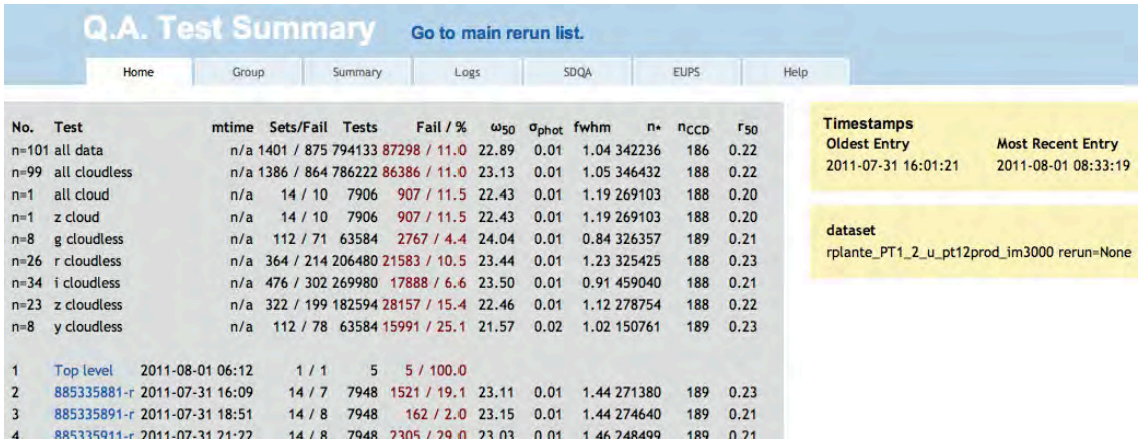

Figure 5-1: Summary page of pipeQA quality results for the Data Release. Click one of the visit IDs (*second column*) to view quality results and plots for a given visit.

Clicking one of the visit identifiers will bring up a page with links to the summary pages of test results. The following types of assessments are currently available:

- Astrometric accuracy
- Completeness of object recovery
- Photometric fidelity (assessed with 8 comparisons among 4 measurements)
- Photometric zero-point
- PSF shape
- Vignetting

Each assessment page provides summary plots and statistics; on the right hand side are one or more graphics of the full FPA, which has active links for exploring the selected assessment report for each

[17] Comparison of results between more than one visit is under development.

CCD (see Figure 5-2). Note that averages over the full FPA are available for some assessments, but in every case the user can display statistics for individual CCDs.

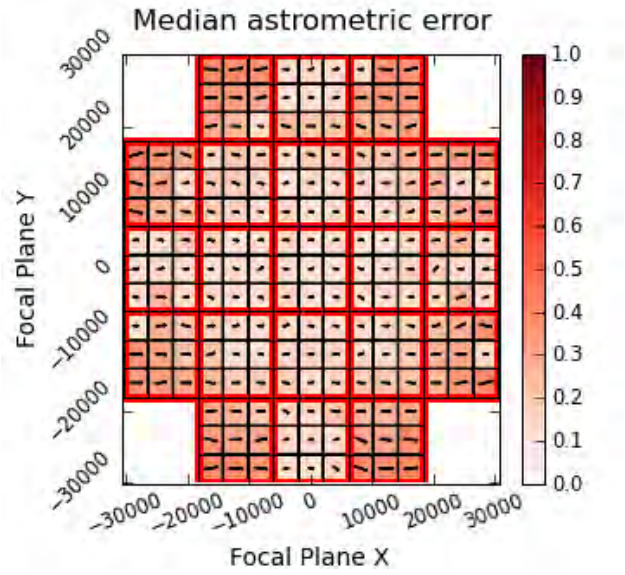

Figure 5-2: Graphic of the full FPA, with color-coded representation for the statistic of interest for each CCD. In this case, the magnitude of the astrometric error is represented (with color code in arcsec at right); the average offset is depicted as a vector inset within each CCD.

Astrometric Accuracy

The accuracy of the world coordinate system (***WCS***) solution can be inferred by comparing the WCS coordinates of bright stars with those of a reference astrometric catalog, which is shown in the summary plots in Figure 5-3. The left figure shows the offset between the measured centroids of matched objects and the catalog position of these objects, represented as a vector field. The top right panel provides the view of these vectors stacked at the position of the reference object, with the green circle representing the radius that contains 50% of the matches. The bottom panel provides a histogram of the offsets, with the median indicated. The width of the offset[18], accumulated over the entire focal plane, is somewhat more than 0.2 arcsec for this visit. The detailed FPA figures for each CCD (not shown) shows the offsets for each matched star.

The RMS of the WCS solution must not be confused with astrometric fidelity in any sense. The RMS is merely a measure of how well the measured centroids of the detected stars match an imperfect coordinate representation: currently a tangent-plane projection plus polynomial distortion correction. The achievable astrometric accuracy requires a global astrometric solution; preliminary analysis suggests this accuracy may be as low as a few to several mas.

[18] The 2-D distribution in this plot would be expected to vary as the Rayleigh distribution: $r \cdot \exp[-r^2 / (2 \cdot \sigma^2)]$, which peaks at $r = \sigma$.

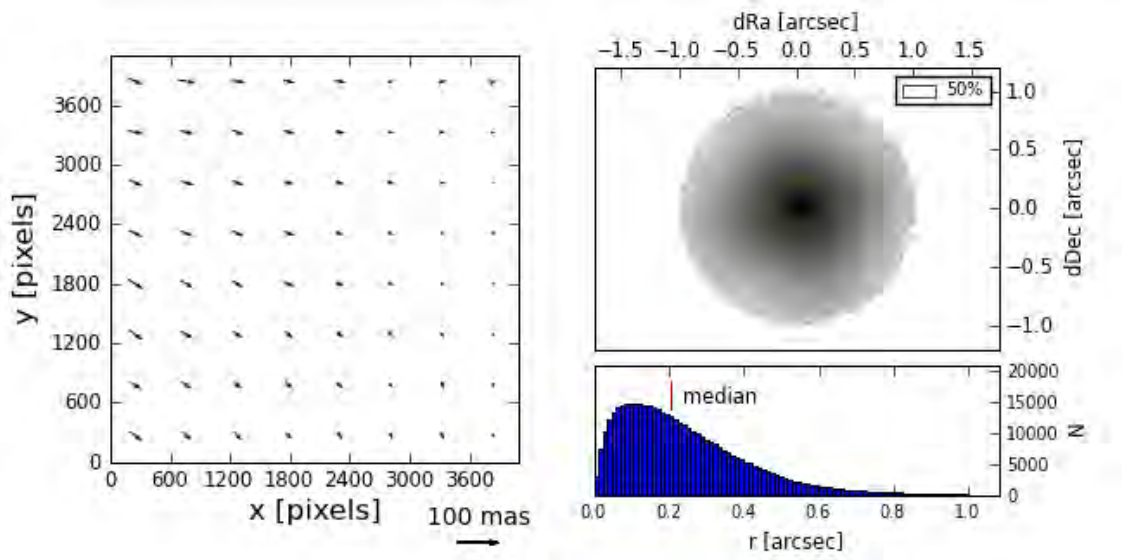

Figure 5-3: Astrometric data quality plots for one exposure in the *r*-band over the CCDs in the FPA. *Left:* Average distribution of offsets between matched star coordinates in the reference catalog and those recovered in pipeline processing. Measured coordinates are determined from the image world-coordinates of the stellar centroids. *Top right*: Stacked offsets from reference catalog position for all stars; *green circle* indicates the radius containing 50% of the matches. *Bottom right*: Histogram of the offsets, with median value indicated.

Object Completeness

For each CCD, the detection completeness for stars is assessed based on the counts of four classes of objects, as a function of magnitude. *Matched* objects are detections that match 1-to-1 with reference catalog; *Blended* objects match N-to-1 with the reference catalog; *Orphan* objects do not match any entry in the reference catalog, and may include false positives and asteroids; while *Unmatched* objects are those present in the reference catalog but not detected in the science image. The measure of completeness is computed as (matched + blended) / (matched + blended + unmatched), indicated with the blue line in the top-left panel of Figure 5-4. The magnitude below which the completeness is < 50% is indicated with the vertical line. The bottom-left panel shows the same breakdown for galaxies only. Orphans are included in both plots. The summary FPA figure provides a visual representation of the photometric depth, as shown in Figure 5-4.

The completeness plots likely paint too pessimistic a picture: they do not account for objects in the outermost 18 pixels of each CCD, which were not searched. This undercounting will be fixed in a future data release. Moving objects are also missing from the reference catalogs, which causes real detections to appear spurious.

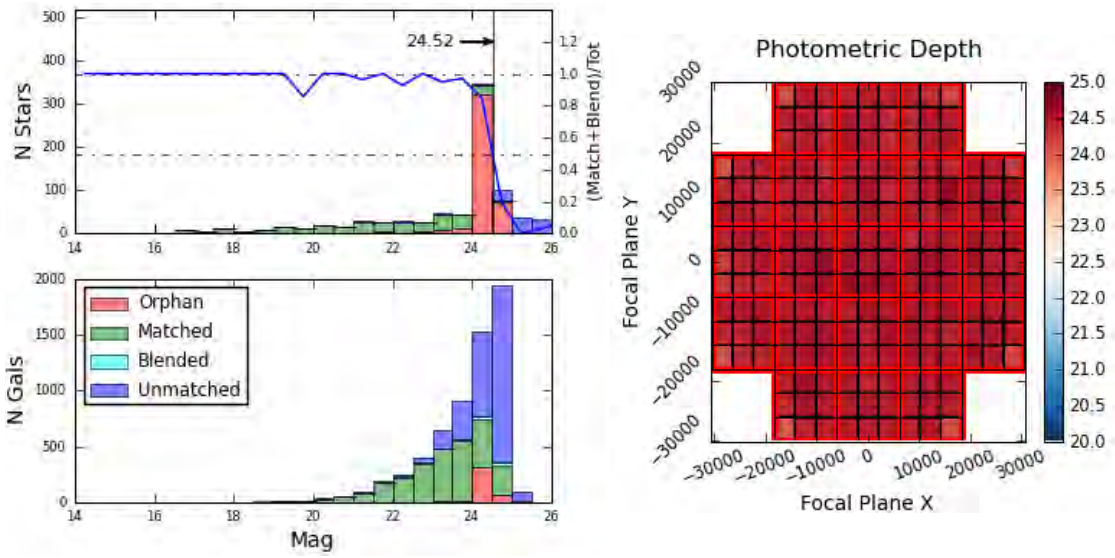

Figure 5-4: Histograms of the detected stars (*upper left*) and galaxies (*lower left*) in the ImSim reference catalog as a function of object magnitude. Objects in this field (Raft 2,2, Sensor 1,1) were *matched* uniquely, matched but *blended* with another object, not matched, or not detected (see text). Summary diagram (*right*) shows photometric depth across the FPA.

Photometric Fidelity

The fidelity and accuracy of the photometry for this data release is determined by comparing the magnitudes as measured in a variety of ways. For each CCD, the diagnostic diagrams show the difference in magnitudes $m_1 - m_2$ as a function of m_1 for the photometric measurements shown in Table 5-2 below (see also Chapter 4.1.4).

Table 5-2: Types of Source Brightness Measurements

Magnitude	Description
m_{CAT}	Source magnitude from the reference "truth" catalog
m_{AP}	Source magnitude from counts within a circular aperture; zero-point set so that, for stars, the aperture flux equals the total flux
m_{INST}	Source magnitude derived from Multi-fit (i.e., multiple component) model
m_{MOD}	Source magnitude derived from Gaussian model
m_{PSF}	Source magnitude derived from PSF size/shape at the position of the source

The pipeQA web pages provide links to five inter-comparisons among pairs of these types. Note that grey extinction from clouds was turned off for some of the simulations, so that measurement and calibration effects can be separated more cleanly from that of simulated weather.

Many of the simulation images for DC3b were produced with grey extinction from clouds turned off. Images with IDs ending in "1" do not include the effects of variable extinction across the FoV, which makes it easier to evaluate the fidelity of single-frame photometric calibration. IDs ending in "0" include the effects of cloud structure in the FoV.

The agreement between cataloged and measured PSF magnitude for all objects (stars + galaxies) shows generally good agreement at brighter magnitudes, as shown in Figure 5-5; the scatter below r~21 is consistent with photon statistics.

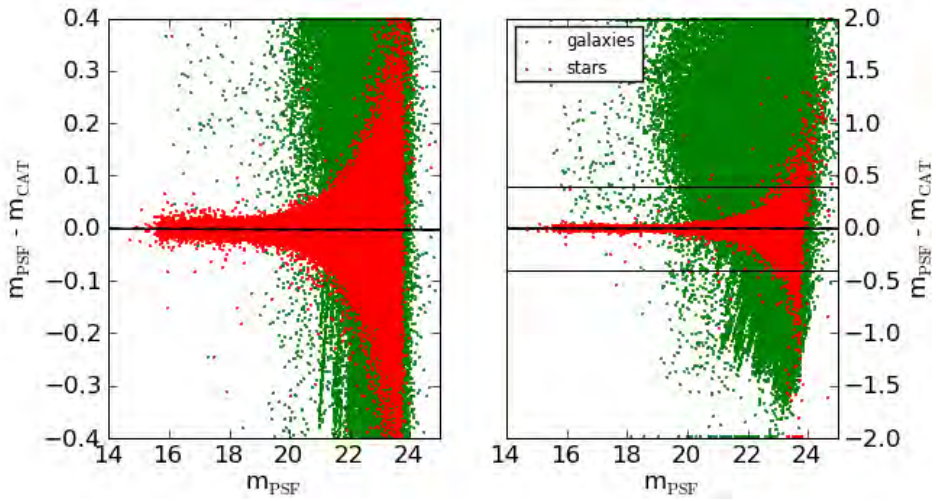

Figure 5-5: Comparison of the PSF magnitudes vs. those in the input ImSim catalog (*right*) and zoomed in (*left*). Comparison is over the full FPA, where stars (*red points*) are distinguished from galaxies (*green*). The dispersion for m_{PSF}>21 to the limiting magnitude near 24 increases as expected for photon statistics.

The color-coded FPA diagrams in Figure 5-6 show the deviation from the expected offset (zero, in this case), slope (also zero), and the standard deviation as a function of magnitude.

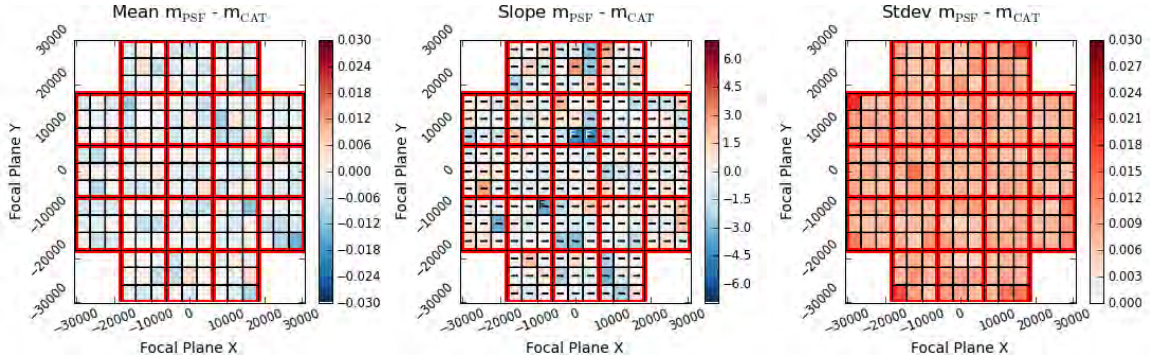

Figure 5-6: FPA diagrams for the comparison in Fig. 5-5, showing the deviation from the expected mean (*right*), slope of the best-fit relation (*center*), and standard deviation of the fit (*left*). CCDs where parameters of the photometric relation exceed permitted thresholds are marked as failed (F).

Clicking on any of the squares in the FPA diagrams shows the results for a single CCD, where sources identified as bright stars are plotted in red. The width of the bright end of this distribution reflects the systematic floor in these measurement comparisons. For PSF magnitudes, this is typically 1-2%. The relationship of magnitude difference as a function of magnitude *for a single CCD* is shown in Figure 5-7 below.

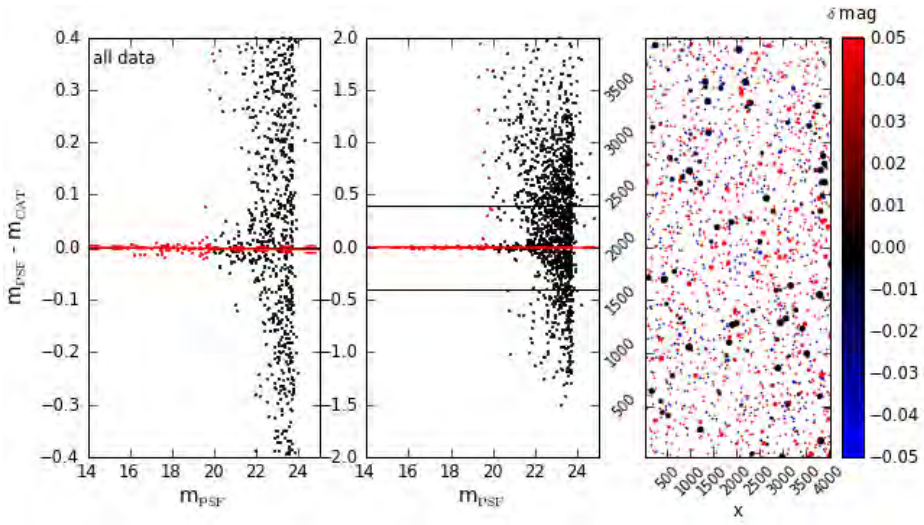

Figure 5-7: Comparison of the PSF magnitudes vs. those in the input ImSim catalog (*center*) and zoomed in (*left*) for a single CCD. Stars (red) define the photometric calibration, and the linear trend is shown (*red dashed line*). Distribution of sources in the CCD is also shown (*right*), color-coded by deviation in mag from the trend.

Photometric Zero-point

For each CCD the central panel in the diagnostic diagram (shown in Figure 5-8) shows the instrumental magnitude of matched stars and galaxies, plotted as a function of the catalog magnitude of the reference objects to which they were matched.

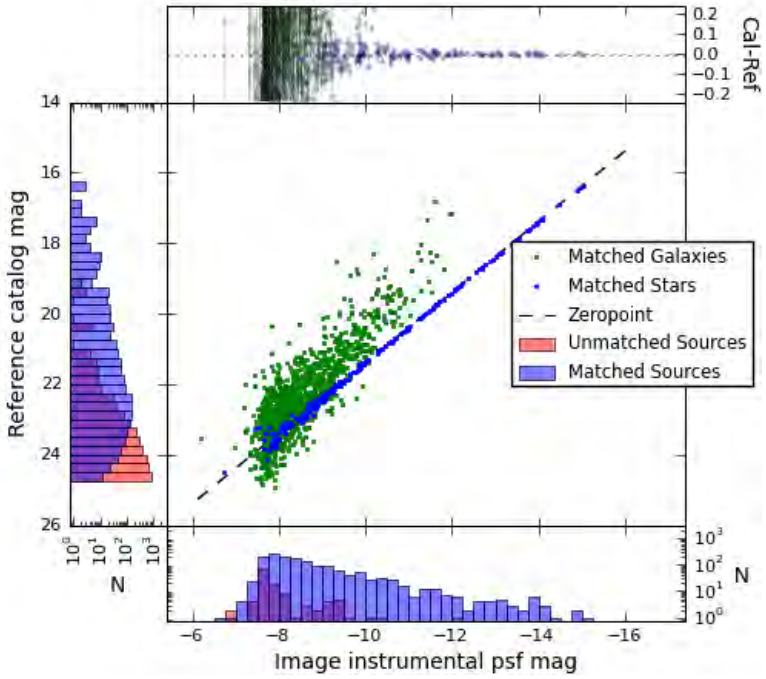

Figure 5-8: Determination of the zero-point magnitude for the central CCD. Star PSF magnitudes show small dispersion but that for galaxies is (as expected) much larger. Some sense of source completeness vs. moving or transient objects and false detections can be inferred from the histograms at left and bottom, but see Figure 5-4 for a detailed comparison.

The fitted relation is shown (*dashed line*). The bottom panel shows a histogram of the uniquely *matched* and *orphaned* sources (detected objects with no entry in the reference catalog) as a function of instrumental magnitude, while the left panel shows a histogram of the matched and unmatched entries in the reference catalog. The top panel shows the scatter of the matched stars and galaxies around the zero-point fit with the median offset of star from the zero-point indicated (*dotted line*). The summary FPA figures (not shown) illustrate the median offset of stars from the zero-point across the focal plane, as well as the fitted zero-point.

PSF Shape

For each CCD, the diagnostic diagram (see Figure 5-9) the ellipticity of star shapes used in the PSF model are plotted as a function of position in the focal plane. The summary FPA figures show the median vector (offset and angle) of the ellipticity for each chip, as well as the effective FWHM in arcsec for the final PSF model.

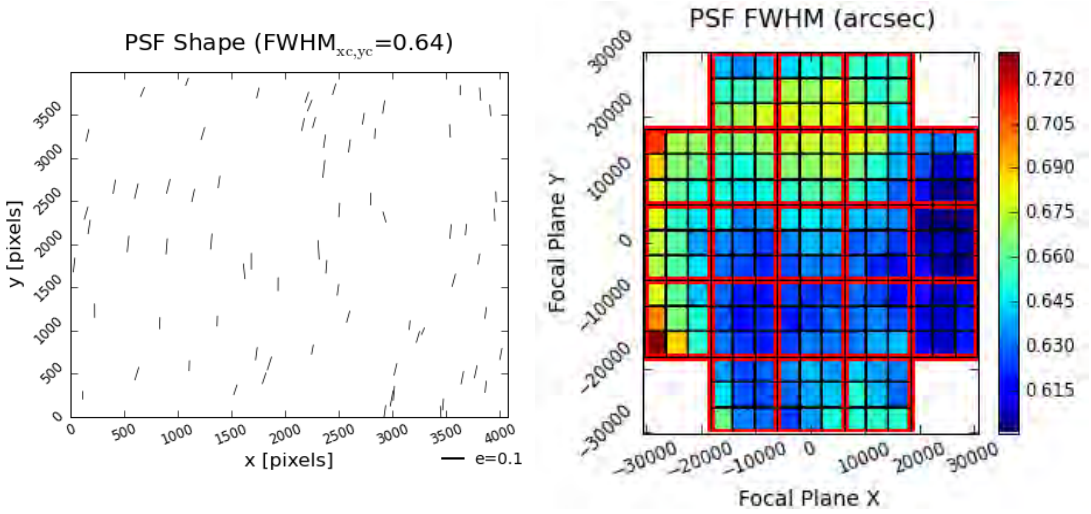

Figure 5-9: Magnitude and orientation of the PSF ellipticity for all stars in the central CCD (*left*). The variation of the model PSF width in arcsec across the FPA (*right*) is also shown. (The variation of ellipticity parameters across the FPA is not shown here.)

Note that the PSF structure and variation across the FPA as generated by ImSim has not yet been validated. Thus, ellipticity patterns such as those in Figure 5-9 have not yet been fully evaluated and may not represent a genuine artifact.

Vignetting

As shown in Figure 5-10, for each CCD the difference between the PSF and reference catalog magnitudes is plotted as a function of radial location from the center of the focal plane. The summary FPA figures show the median offset, as well as the standard deviation of this offset (not shown), for each chip. The edges of the FPA are the most likely to show poor photometric fidelity when the vignetting is not adequately corrected. The cause of this problem is believed to be an artifact of the ImSims. Specifically, the flat-fields and sky background are not generated in the same way as the photons for the astrophysical sources. This is dealt with by creating an empirical correction for the vignetting, and applying it in the pipeline.

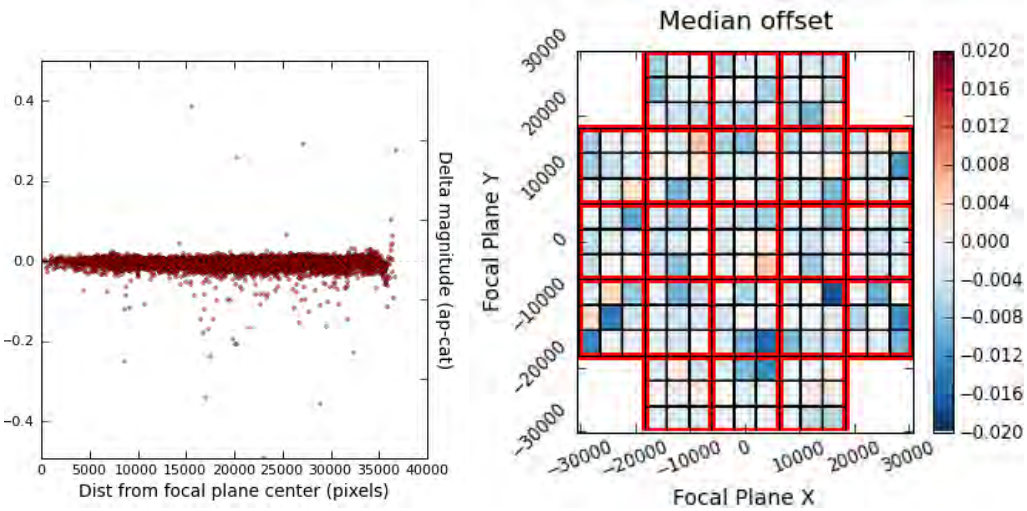

Figure 5-10: Difference of the *r*-band PSF magnitudes with those in the reference catalog vs. distance from the center of the FPA (*left*). This summary plot is a stack of the results from all CCDs in the FPA. Offsets for individual CCDs can be selected from an FPA graphic (*right*). The standard deviation about the offset (not shown) is also a sensitive indicator of the fidelity of the vignetting correction.

The correction for vignetting at the edge of the focal plane was derived for ImSim images by computing residuals (stellar PSF mag – reference mag) as a function of radius from a single *r*-band image. This provides a measurement of the multiplicative correction to the measured flux necessary to bring the measurements in line with the reference. In practice, the flux residuals were binned in radius, and the median in each radial bin contributed to a spline fit to the correction as a function of radius. Full chip correction images were produced using the correction function assuming azimuthal symmetry. The *r*-band correction image was used for all bands since there was not enough data in other bands to constrain the correction function.

The flat frames were simulated using an uncorrected analytic vignetting function. The sky background was applied to the simulated images using the same uncorrected vignetting function. Thus the flats were applied so that the background can be fit, and the flux correction was applied (as a multiplication) after background subtraction but before measurement. The photometric errors show that the correction was effective: the typical residual error is <2%. In the extreme corners of the focal plane these residuals can still climb to 5% or so where the correction is rising most rapidly and the data constrain the correction least well.

The formulation of the vignetting correction borrows from prior knowledge of the Image Simulation algorithms and their (imperfect) match to the modeled LSST optical system. This is different from the empirical approach that will be taken during operations, which will likely derive the vignetting function from the global photometric solution.

5.2. Other Assessments

5.2.1. Input Image Simulation Data

As described in Chapter 3, the ImSim data are derived from a relatively high-fidelity computational model of the expected image properties. But at present not all expected properties are included in either the input catalog or the detector physics. It is likely that some of these missing properties will be important for determining whether one or more science goals can be achieved with LSST. Examples include: extended objects such as structure in galaxy profiles, galaxy halos and streams, Galactic nebulae; extended variable objects such as symbiotic nebulae and supernova light echos; etc.

5.2.2. Processing Algorithms

The algorithms and processes used in the processing pipelines have been described in Chapter 4 in enough detail to convey a basic understanding of how raw data have been reduced and calibrated. For the present, a more detailed understanding would require understanding the source code and software configuration (which is available publically), which is a tall order for most people. However, evaluating the effectiveness of the algorithms can be accomplished in a number of direct and indirect ways. ImSim image headers, for instance, contain the location and orientation of the artificial cosmic rays so it is possible to compare the CR-flagged pixels directly with the simulation parameters. CFHT-LS data have been independently (and very carefully) processed with the Terapix pipeline, which would be an excellent yardstick with which to measure the effectiveness of DC3b processing. More generally, independent catalogs, software or user scripts can be very useful for generating measurements such as source fluxes, shapes, and locations for comparison with those recorded in the source catalog.

5.2.3. Advanced Data Quality Assessments

With the pipeQA results as a baseline, one may imagine a number of ways to characterize the data and the science quality. A very incomplete list includes:

- How does the photometric depth vary with PSF size (which is a function of the seeing), or with background level?
- How does the quality of the star/galaxy separation degrade with seeing, or image background level?
- To what extent does crowding (i.e., blended sources) affect the star/galaxy separation?
- Are there ways to more finely tune the source association to reduce blends? Can the tuning include seeing as a parameter?

5.2.4. Output Data Products

The Data Release production will generate most of the kinds of data products that are planned to support community science during LSST operations. Over time, the DM team will provide tools for advanced search, access, and analysis of archived data products. The Data Challenges, particularly DC3b, give the Science Collaboration community the opportunity to review the form, format, organization, and content of these products and assess their suitability to enable the key science drivers. Areas where feedback would be particularly useful include:

1. Are there other data products needed, beyond those delivered (and planned: see Chapter 2) for DC3b?
2. Data product content: are other kinds of concomitant data (beyond quality masks and variance planes) needed?

3. In a future release, it is likely that calibrated images will be offered in a compressed format, such as that proposed in the FITS tiled image compression convention[19] that is seeing wider use in the community. Would a slightly lossy compression option (i.e., coarser-grained sampling of the noise) present a problem for your science?
4. Are current services (see Chapter 2) adequate for searching, querying, and retrieving data products (both catalog data and images)? Future plans call for direct and programmatic access to (a copy of) the science database, with support for SQL. What sorts of support and services are desired in this context?

It may well be that some kinds of advanced data processing, including data subsetting, for certain scientific purposes do not fit within the LSST project scope. In these cases the Science Collaborations may find it to their advantage to create tools, services, and data products (the so-called *Level 3* data products) that serve their science needs, and starting sooner on that development would be an advantage.

5.3. Community Feedback

One of the major objectives of publishing data from the DC3b data release production is for a wider, scientific audience to participate in the evaluation of the data, and to bring more expertise to bear on the technical challenges of data reduction and analysis. Feedback can take a number of forms, and will cover a wide variety of topics.

5.3.1. General Feedback

It is perhaps useful to summarize the types of feedback, what form it might usefully take, and the mechanisms that are in place for receiving it. The following are meant as guidelines.

This Handbook. If you find errors, incomplete information, or have requests and suggestions for improving this *Handbook*, please contact the HelpDesk: dc-help@lsst.org.

Science/Technical Issues. Please post to the Science User Forum (https://www.lsstcorp.org/sciencewiki/index.php?title=Special:AWCforum) any suggestions you have for new algorithms, approaches, and techniques for either analyzing the science quality of the data, or for improving the pipeline processing. This forum is also a good place to engage in discussions of what analyses might be most important or fruitful, and to see what other Collaboration members are doing.

Results of Analysis. If you have conducted some analysis of DC3b data products and have a written analysis in hand, these are likely best posted on the Science Wiki: http://www.lsstcorp.org/sciencewiki/index.php?title=Main_Page. Naturally the more thoroughly documented the analysis, the more useful it will be to others, including the DM team. In particular, it is important to describe which data were analyzed (visit IDs for images, selection criteria for catalogs).

Questions about Pipeline Processing: Detailed questions about how the pipelines work should be directed to the HelpDesk: dc-help@lsst.org.

Technical problems. If users have problems with accessing data products, the access tools (**Gator** and VOInventory), or if the related services do not respond or do not provide sensible results, please report this to the HelpDesk: dc-help@lsst.org. Please be as specific as possible about the date/time when the problem occurred, the input you provided, and any error messages that resulted.

[19] See http://fits.gsfc.nasa.gov/registry/tilecompression.html.

5.3.2. Science/Technical Feedback

Evaluation of science data quality for LSST will be an involved and sometimes complex process, and it is clear that the Data Management design effort will for the near term be limited to relatively targeted, mostly basic assessments. The following is a partial list of scientific or technical analyses that, were they completed, would contribute significantly to the assessment of the data products and software in DC3b. This is an area where the Science Collaboration members can contribute substantially to the Data Management development effort.

1. Assess whether all pixels in ImSim images that were identified as being affected by cosmic rays correspond to the list of artificial cosmic rays that were introduced in the model.
2. Evaluate the fidelity of the ImSim images, and identify important phenomena that are not yet modeled, or shortcomings with the current model.
3. Evaluate the veracity and fidelity of source identification. Will tuning the current algorithm suffice, or is another algorithm needed?
4. Evaluate the effectiveness of the source association algorithm. For the present, objects that associate to unique sources for the most part arise from spurious sources. Is there a way to tune the algorithm to recover genuine, unique sources (which of course will constitute a critical class of LSST science)? Would another algorithm work better?
5. Evaluate the data products themselves (described in the previous subsection).

And finally, it would be most helpful to receive additions to the above list of analyses that would help characterize the data quality, and to identify problems and shortcomings where they exist. Please post these ideas on the Science User Forum at https://www.lsstcorp.org/sciencewiki/index.php?title=Special:AWCforum.

5.4. References and Further Information

Contributing Authors

The following individuals contributed to the material presented in this chapter: Tim Axelrod, Andy Becker, Steve Bickerton, Simon Krughoff, Dustin Lang, Robert Lupton, and Dick Shaw.

References

Mellier, Y., et al. 2008, *The CFHTLS T005 Release*, (Paris: Institut d'Astrophysique de Paris). Available at: http://terapix.iap.fr/cplt/oldSite/Descart/CFHTLS-T0005-Release.pdf

Shaw, R. A., Levine, D., Axelrod, T. S., Laher, R. R., & Mannings, V. G. 2010, *Science Data Quality Assessment for the LSST*, Proc. SPIE, 7740, 14

Glossary

The following glossary of technical terms that were used in this *Handbook* includes many of those that can be found on the Data Management Wiki.

Term	Definition
Adaptive (Gaussian) moments	The second moments of the source intensity distribution, measured using a scheme designed to have near-optimal signal-to-noise ratio.
Alert	Refers to the structured communication that is issued rapidly via the internet to the community that characterizes detection of one or more sources that are new, or have changed significantly in position or brightness, relative to the applicable ***image template***.
Amplifier	Electronic component of a ***CCD*** that is used to recover the signal during read-out. For LSST, multiple amplifiers on each CCD will enable simultaneous read out of adjacent regions of the detector. Often this term is used as a synonym for a read-out ***channel***.
Calibrated image	An image from a single visit to a region of sky that is the combination of two ***snap images***, each of which has been corrected for instrumental signature. The ***WCS*** determination, photometric calibration, and various other characterizations have also been determined for calibrated images. Calibrated images have an effective exposure time equal to the sum of the components, which nominally is 30 s.
CCD	Charge-coupled device. This is the type of ***sensor*** that will be used in the LSST camera for detecting and recording radiation in the visible band. Contiguous portions of a CCD detector can be read out simultaneously through parallel output ***channels*** if the electronic design includes multiple ***amplifiers***.
CFHT-LS	Five-band legacy imaging survey conducted at the Canada-France-Hawaii Telescope from 2003—2008.
Channel	The output from a specific ***amplifier*** (one of many) from a single ***sensor***. Often used as a synonym for amplifier, in the sense of referring to a specific region of a ***CCD***.
Data Challenge	A structured set of activities that processes large volumes of astronomical data using the Data Management software stack over a massively parallel, high-throughput computing platform. The data challenges aim, over time, to demonstrate the ability to produce images and catalogs with the scientific fidelity, scalability, and throughput comparable to that expected from LSST.
DIA	Difference Image Analysis. Refers to the data products or catalog entries that are generated during the pipeline stage by this name, which include the detection new ***sources*** and brightness changes in known ***objects***.
Difference image	A pixel-by-pixel difference between the image being processed and an ***image template***, where the template has been warped to the same geometry, photometrically scaled, and background-matched. The resulting difference image is zero everywhere, apart from shot noise and objects that are new, or have changed in brightness or position, relative to the template.

Term	Definition
Exposure	One of a pair of ***raw images*** in a single band, taken sequentially, of the same area of sky. Pairs of exposures facilitate the identification of cosmic ray artifacts. This term is a synonym for the less commonly used term ***snap image***.
FITS	Flexible Image Transport System, which is the astronomical (IAU) standard for structured data in a file. Allowed contents include images, tables, and ***metadata*** stored in ASCII headers.
Footprint	The spatial extent of an image or of an astronomical object within an image. The footprint can be irregular in shape, such as that of a galaxy, or of a group of spatially overlapping but non-coincident images.
FPA	Focal Plane Array, or a regular grid of ***sensors*** placed at the focus of the imaging camera.
FoV	Field of view, used to describe the spatial extent of the sky observed with the ***FPA*** or with a portion of it (e.g., a single detector).
HDU	Header and Data Unit, or a data structure in a ***FITS*** file that consists of an ASCII header and the data that the header describes. Note that a (primary) HDU may consist only of a header, with no data blocks. FITS extensions are structured as HDUs that appear after the primary HDU in a FITS file.
Image Extension	An ***HDU*** within a ***FITS*** file consisting of a header plus a binary image array. The pixel values may be expressed in any form allowed by the FITS Standard (e.g., integer or floating-point).
ImSim	Can refer either to the high fidelity LSST image simulation program, or to the simulated LSST camera images that the ImSim software generates.
Image Template	An image of a section of sky in a single band that is deep, of very high image quality, and where all transients, moving objects, and artifacts have been removed. Such images are used as templates to perform ***difference image*** analysis in order to detect variable, transient, and moving objects.
Metadata	Strictly speaking, information or data that describe other data, such as an image. Most metadata are stored in the ***science database***. Metadata also appear in the keyword/value pairs in the headers of FITS files.
MSS	Mass storage system, which stores large volumes of data on a variety of media (including spinning disk and tape), whose contents appear to a user to be a regular file system.
Object	Refers to an *astronomical object*, such as a star, galaxy, asteroid, or other physical entity. Objects can be static, or change brightness or position with time. Generally an object will be associated with more than one instance of a ***source*** detection.
Pipeline	A unit of data processing software that is independently executable within the Data Management System, and which performs a logically connected sequence of operations. Pipelines are composed of one or more processing ***stages***.
Production	Episodes of data processing that are organized to achieve a particular purpose, such as issuing event ***alerts*** during a night's observing; generating ***calibrated images***, ***source*** and ***object*** catalogs, and ***image templates*** on an annual basis; or constructing calibration reference products.
Provenance	The structured description of the origin of a data product, including its processing history, data dependencies, and software version identifier. All provenance information is stored in the science database, although some of it is replicated in the headers of image data products.

Term	Definition
PSF	Point Spread Function, or the two-dimensional brightness profile of a point source (i.e., an unresolved astronomical object) as it is realized by the detector, including all effects of the atmosphere, telescope/camera optics, and detector sampling.
Raft	Physical sub-structure in the ***FPA*** to which a 3 × 3 grid of ***CCD*** detectors is mounted. See Figure 3-1.
Raw Image	A regular array of pixel data, and associated ***metadata***, that were obtained from the observing environment in a single exposure from one or more detectors in the ***FPA***.
Science Database	The repository of all metadata that describe all observations, their provenance, and their quality attributes; and of the measurements that have been made on the images, such as source positions, brightnesses, and other attributes.
SDQA	Science Data Quality Analysis. The software and processes that measure quality items of interest (metrics), and compare them to thresholds that define nominal conditions.
Sensor	Generic engineering term used to refer to a single detector, which in the case of the LSST is a ***CCD***s.
Sky Tile	A region of sky with an extent that depends upon position and sky tiling scheme. For many regions the extent is roughly 0°.5 × 0°.5. Full-sky images, such as templates, deep-detection co-adds, etc. will be partitioned in this way.
Snap Image	One of a pair of ***raw images*** in a single band, taken sequentially, of the same area of sky. Pairs of snaps facilitate the identification of cosmic ray artifacts. This term is a synonym for the more commonly used term ***exposure***.
Source	A single detection of an astrophysical ***object*** in an image, the characteristics for which are stored in the Source Catalog of the science database. The Data Management System attempts to associate multiple source detections to single ***objects***, which may vary in brightness or position over time.
Stage	A portion of a ***pipeline*** that performs a discrete algorithmic operation, and which is not independently executable.
Tracklet	Trial apparent path of a moving object, defined by the positions of two source detections that are thought to be related. Tracklets must be further associated with other tracklets in order to determine an orbit for a solar system ***object***.
VAO	Virtual Astronomical Observatory (formerly NVO). The organization in the United States that provides software and services for discovering, exploring, analyzing, and publishing federated, digital astronomical data resources on the internet.
Visit	A set of (usually two) exposures, taken sequentially with the same filter at the same position on the sky. Multiple images are used in the productions to screen for transient artifacts such as cosmic rays and satellite trails.
WCS	World Coordinate System, which is the specification of a mapping between detector coordinates (i.e., pixels) to a reference system such as celestial coordinates. The formalism for the WCS mapping is defined in the ***FITS*** standard.