Large Synoptic Survey Telescope
Data Products Definition Document
(*** DRAFT ***)

Mario Jurić <mjuric@lsst.org>

with input from

T. Axelrod, A.C. Becker, J. Becla, G.P. Dubois-Felsmann, M. Freemon, Ž. Ivezić, J. Kantor, K-T Lim, R. H. Lupton, D. Shaw, M. Strauss, *and* J.A. Tyson

for the LSST Project

May 22, 2013

Abstract

This document describes the plans for contents of Level 1 and 2 LSST data products, and the rationale behind various choices that were made. This is an **internal draft** and a work in progress. **It should not be circulated widely until this notice is removed.**

Contents

1 Introduction

LSST will be a large, wide-field ground-based optical telescope system designed to obtain multiple images covering the sky that is visible from Cerro Pachón in Northern Chile. The current baseline design, with an 8.4m (6.7m effective) primary mirror, a 9.6 deg^2 field of view, and a 3.2 Gigapixel camera, will allow about 10,000 square degrees of sky to be covered using pairs of 15-second exposures twice per night every three nights on average, with typical 5σ depth for point sources of $r \sim 24.5$ (AB). The system is designed to yield high image quality as well as superb astrometric and photometric accuracy. The total survey area will include 30,000 deg^2 with $\delta < +34.5°$, and will be imaged multiple times in six bands, *ugrizy*, covering the wavelength range 320–1050 nm. The project is scheduled to begin the regular survey operations at the start of next decade. About 90% of the observing time will be devoted to a deep-wide-fast survey mode which will uniformly observe a 18,000 deg^2 region about 1000 times (summed over all six bands) during the anticipated 10 years of operations, and yield a coadded map to $r \sim 27.5$. These data will result in databases including 10 billion galaxies and a similar number of stars, and will serve the majority of the primary science programs. The remaining 10% of the observing time will be allocated to special projects such as a Very Deep and Fast time domain survey.

The LSST will be operated in fully automated survey mode. The images acquired by the LSST Camera will be processed by LSST Data Management software to a) detect and characterize imaged astrophysical sources and b) detect and characterize changes in time in LSST-observed universe. The results of that processing will be reduced images, catalogs of detected objects and the measurements of their properties, and prompt alerts to "events" –

changes in astrophysical scenery discovered by differencing incoming images against older, deeper, images of the sky in the same direction (*templates*, see §2.4.3).

The *broad*, *high-level*, requirements for LSST Data Products are given by the LSST Science Requirements Document. This document lays out the *specifics* of what the data products will comprise of, how those data will be generated, and when. It serves to inform the flow-down from the *LSST Science Requirements Document* and the *LSST Observatory System Specifications*, to the *LSST Data Management System Requirements* document, the UML model, and the database schema.

1.1 Level 1 and 2 Data Products

LSST Data Management will perform two, somewhat overlapping in scientific intent, types of image analyses:

1. Analysis of difference images, with the goal of detecting and characterizing astrophysical phenomena revealed by their time-dependent nature. The detection of supernovae superimposed on bright extended galaxies is an example of this analysis. The processing is done on a nightly or daily basis and produces **Level 1** data products. They include the difference images, the sources detected in difference images (`DIASources`), astrophysical objects[1] these are associated to (`DIAObjects`), and Solar System objects (`SSObjects`[2]). These are added to the **Level 1 database** and made available in real time. Notifications ("alerts") about new `DIASources` will be issued using community-accepted standards within 60 seconds of observation.

2. Analysis of direct images, with the goal of detecting and characterizing astrophysical objects. Detection of faint galaxies on deep coadds and their subsequent characterization is an example of this analysis. The

[1]The LSST has adopted the nomenclature by which single-epoch detections of astrophysical *objects* are called *sources*. This nomenclature is not universal: some surveys call *detections* what we call *sources*, and use the term *sources* for what we call *objects*.

[2]`SSObjects` used to be called call "Moving Objects" in previous versions of the Data Products baseline. The name is potentially confusing as high-proper motion stars are moving objects as well. A more accurate distinction is the one between objects inside and outside of the Solar System.

results are **Level 2** data products. These products, released annually[3], will include the single-epoch images, deep coadds, catalogs of `Objects` (detections on deep coadds) and `Sources`[4] (measurements on individual direct images), as well as fully reprocessed Level 1 data products (see §2.3.7). In contrast to the Level 1 database, which is updated in real-time, the Level 2 databases are static and will not change after release.

The two types of analyses have different requirements on timeliness. Changes in flux or position of objects may need to be immediately followed up, lest interesting information be lost. Thus the primary results of analysis of difference images – discovered and characterized `DIASources` – generally need to be broadcast as *event alerts* within 60 seconds of end of visit[5] acquisition. The analysis of science (direct) images is less time sensitive, and will be done as a part of annual data release process.

2 Level 1 Data Products

2.1 Overview

Level 1 data products are a result of difference image analysis (DIA; §2.2.1). They include the sources detected in difference images (`DIASources`), astrophysical objects that these are associated to (`DIAObjects`), identified Solar System objects[6] (`SSObject`), and related, broadly defined, metadata (including eg., cut-outs[7]).

`DIASources` are sources detected on difference images (those above $S/N = 5$ after correlation with the PSF profile). They represent changes in flux with respect to a deep template. Physically, a `DIASource` may be an observation

[3]Except for the first two data releases, which will be created six months apart.

[4]When written in bold monospace type (i.e., `\tt`), `Objects` and `Sources` refer to objects and sources detected and measured as a part of Level 2 processing.

[5]The LSST takes two (nominally 15 second) exposures per pointing, called *snaps*. That pair of exposures is called a *visit*.

[6]The LSST SRD considers Solar System object orbit catalog to be a Level 2 data product (LSSTSRD, Sec 3.5). Nevertheless, to successfully differentiate between apparitions of known Solar System objects and other types `DIASources` we consider it functionally a part of Level 1.

[7]Small, 30×30, sub-images at the position of a detected source. Also known as *postage stamps*.

of new astrophysical object that was not present at that position in the template image (for example, an asteroid), or an observation of flux change in an existing source (for example, a variable star). Their flux can be negative (eg., if a source present in the template image reduced its brightness, or moved away).

Clusters of `DIASources` detected on visits taken at different times are associated with either a `DIAObject` or an `SSObject` to represent the underlying astrophysical phenomenon. The association can be made in two different ways: by assuming the underlying phenomenon is an object within the Solar System moving on an orbit around the Sun[8], or by assuming it to be distant enough to only exhibit small parallactic and proper motion[9]. The latter type of association is performed during difference image analysis right after the image has been acquired. The former is done at daytime by the Moving Objects Processing Software (`MOPS`), unless the `DIASource` is an apparition of an already known `SSObject`. In that case, it will be flagged as such during difference image analysis.

At the end of the difference image analysis, we will generate alerts for all newly discovered `DIASources`.[10].

2.2 Level 1 Data Processing

2.2.1 Difference Image Analysis

The following is a high-level description of steps which will occur during regular difference image analysis:

1. A visit is acquired and reduced to a single *visit image* (cosmic ray rejection, instrumental signature removal[11], combining of snaps, etc.).

[8] We don't plan to fit for motion around other Solar System bodies; eg., identifying new satellites of Jupiter is left to the community.

[9] Where 'small' is small enough to unambiguously positionally associate together individual apparitions of the object.

[10] For observations on the ecliptic near the opposition, Solar System objects will dominate the `DIASource` counts, and (until they're recognized as such) overwhelm the explosive transient signal. It will therefore be advantageous to quickly identify the majority of Solar System objects early in the survey.

[11] Eg., subtraction of bias and dark frames, flat fielding, bad pixel/column interpolation, etc.

2. The visit image is differenced against the appropriate template and `DIASources` are detected.

3. The flux and shape[12] of the DIASource are measured on the difference image. The visit image is force-photometered at the position of the `DIASource` to obtain a measure of the absolute flux. No deblending will be attempted.

4. The Level 1 database (see §2.3) is searched for a `DIAObject` or `SSObject` with which to positionally associate the observed `DIASource`[13]. If no match is found, a new `DIAObject` is created and the observed `DIASource` is associated to it.

5. If the `DIASource` has been associated with an `SSObject` (a known Solar System object), it will be flagged as such and an alert will be issued. Further processing will occur in daytime (see section 2.2.2).

6. Otherwise, the associated `DIAObject` measurements will be updated with new data. All affected columns will be recomputed, including proper motions, centroids, light curves, etc.

7. The Level 2 database[14] is searched for one or more `Objects` positionally close to the `DIAObject`, out to some maximum radius[15]. The IDs of these `Objects` are recorded in the `DIAObject` record and provided in the event alert.

8. An alert is issued that includes: the name of the Level 1 database, the timestamp of when this database has been queried to issue this alert, the `DIASource` ID, the `DIAObject` ID[16], name of the Level 2 database and the IDs of nearby `Objects`, and the associated science

[12]The "shape" in this context consists of weighted 2$^{\text{nd}}$ moments, as well as a fit to a trailed source model.

[13]The association algorithm will guarantee that a `DIASource` is associated with not more than one `DIAObject` or `SSObject`. The algorithm will take into account the parallax and proper or Keplerian motions, as well as the errors in estimated positions of `DIAObject`, `SSObject`, and `DIASource` to find the maximally likely match. Multiple `DIASources` in the same visit will not be matched to the same `DIAObject`.

[14]Level 2 database is a database resulting from annual data release processing.

[15]Eg., a few arcseconds.

[16]We guarantee that a receiver will always be able to regenerate the alert contents at any later date using the included timestamps and metadata (IDs and database names).

content (centroid, fluxes, low-order lightcurve moments, periods, etc.), *including the full light curves.* See Section 2.5 for a more complete enumeration.

9. For all `DIAObjects` overlapping the field of view, to which a `DIASource` from this visit has not been associated, forced photometry will be performed (point source photometry only). Those measurements will be stored as appropriately flagged `DIASources`[17]. No alerts will be issued for these `DIASources`.

10. Within 24 hours of discovery, *precovery* PSF forced photometry will be performed on any difference image overlapping the position of new `DIAObjects` taken within the past 30 days, and added to the database. Alerts will not be issued with precovery photometry information.

In addition to the processing described above, a smaller sample of sources detected on difference images *below* the nominal $S/N = 5$ threshold will be measured and stored, in order to enable monitoring of difference image analysis quality.

Also, the system will have the ability to measure and alert on a limited[18] number of sources detected below the nominal threshold for which additional criteria are satisfied. For example, a $S/N = 3$ source detection near a gravitational keyhole may be highly significant in assessing the danger posed by a potentially hazadous asteroid. The project will define the initial set of criteria by the start of Operations.

2.2.2 Solar System Object Processing

The following will occur during regular Solar System object processing (in daytime[19], after a night of observing):

[17]For the purposes of this document, we're treating the `DIASources` generated by precovery measurements to be the same as `DIASources` detected in difference images (but flagged appropriately). In the logical schema, these may be divided into two separate tables.

[18]It will be sized for no less than $\sim 10\%$ of average `DIASource` per visit rate.

[19]Note that there *is no guarantee on when daytime Solar System processing must finish*, just that, averaged over some reasonable timescale (eg., a month), a night's worth of observing is processed within 24 hours. Nights rich in moving objects may take longer to process, while nights with less will finish more quickly. In other words, the requirement is on *throughput*, not latency.

1. The orbits/physical properties of `SSObjects` that were re-observed on the previous night are recomputed. Updated data are entered to the `SSObjects` table.

2. All `DIASources` detected on the previous night, that have not been matched with high probability to a known `Object`, `SSObject`, or an artifact, are analyzed for potential pairs, forming *tracklets*.

3. The collection of tracklets collected over the past 30 days is searched for subsets forming *tracks* consistent with being on the same Keplerian orbit around the Sun.

4. For those that are, an orbit is fitted and a new `SSObject` table entry created. `DIASource` records are updated to point to the new `SSObject` record. `DIAObjects` "orphaned" by this unlinking are deleted.[20].

5. Precovery linking is attempted for all `SSObjects` whose orbits were updated in this process. Where successful, `SSObjects` (orbits) are updated as needed.

2.3 The Level 1 database

The described alert processing design presupposes the existence of a Level 1 database that contains the objects and sources detected on difference images. At the very least[21], this database will have tables of `DIASources`, `DIAObjects`, and `SSObjects`, populated in the course of difference image and Solar System object processing[22]. As these get updated and added to, their updated contents becomes visible (queryable) immediately[23].

Note that *this database is only loosely coupled to the Level 2 database*. All of the coupling is through providing positional matches between the `DIAObjects` entries in the Level 1 database and the `Objects` in the Level 2 database database. There is no direct `DIASource`-to-`Object` match. The

[20]Some `DIAObjects` may only be left with forced photometry measurements at their location (since all `DIAObjects` are force-photometered on previous and subsequent visits); these will be kept but flagged as such.

[21]It will also contain exposure and visit metadata, MOPS-specific tables, etc. These are either standard/uncontroversial, or implementation-dependent, irrelevant for science, and therefore not discussed here.

[22]The latter is also colloquially known as *DayMOPS*.

[23]No later than the moment of issuance of any event alert that may refer to it.

adopted data model emphasizes that *having a `DIASource` be positionally coincident with an `Object` does not imply it is physically related to it.* Absent other information, the least presumptuous data model relationship is one of *positional association*, not *physical identity.*

This may seem odd at first: for example, in a simple case of a variable star, matching individual `DIASources` to `Objects` is exactly what an astronomer would want. That approach, however, fails in the following scenarios:

- *A supernova in a galaxy.* The matched object in the `Object` table will be the galaxy, which is a distinct astrophysical object. We want to keep the information related to the supernova (eg., colors, the light curve) separate from those measurements for the galaxy.
- *An asteroid occulting a star.* If associated with the star on first apparition, the association would need to be dissolved when the source is recognized as an asteroid (perhaps even as early as a day later).
- *A supernova on top of a pair of blended galaxies.* It is not clear in general to which galaxy this `DIASource` would belong. That in itself is a research question.

`DIASource`-to-`Object` matches can still be emulated via a three-step link (`DIASource-DIAObject-Object`). For ease of use, views or pre-built table with these will be offered to end-users.

In the sections to follow, we present the *conceptual schemas* for the most important Level 1 database tables. These convey *what* data will be recorded in each table, rather than the details of *how*. For example, columns whose type is an array (eg., `radec`) may be expanded to one table column per element of the array (eg., `ra`, `decl`) once this schema is translated to SQL. Secondly, the tables to be presented are normalizes (i.e., contain no redundant information). For example, since the band of observation can be found by joining a `DIASource` table to the table with exposure metadata, there's no column for 'band' in the `DIASource` table. In the as-built database, the views presented to the users will be appropriately denormalized for ease of use.

2.3.1 DIASource Table

This is a table of sources detected at $SNR \geq 5$ on difference images (DIASources). On average, we expect ~ 2000 DIASources per visit ($\sim$ 2M per night; 20,000 per deg^2 per hour).

Some $SNR \geq 5$ sources will not be caused by observed astrophysical phenomena, but by artifacts (bad columns, diffraction spikes, etc.). The difference image analysis software will attempt to identify and flag these as such.

Unless noted otherwise, all DIASource quantities (fluxes, centroids, etc.) are measured on the difference image.

Table 1: DIASource Table

Name	Type	Unit	Description
diaSourceId	uint64		Unique source identifier
ccdVisitId	uint64		ID of CCD and visit where this source was measured
diaObjectId	uint64		ID of the DIAObject this source was associated with, if any.
ssObjectId	uint64		ID of the SSObject this source has been linked to, if any.
midPointTai	double	time	Time of mid-exposure for this DIASource.
radec	double[2]	degrees	(α, δ)[24]
radecCov	float[3]	various	radec covariance matrix
xy	float[2]	pixels	Column and row of the centroid.
xyCov	float[3]	various	Centroid covariance matrix
SNR	float		The signal-to-noise ratio at which this source was detected in the difference image.[25]

Continued on next page

[24] The astrometric reference frame will be chosen closer to start of operations.

[25] This is not necessarily the same as psFlux/psFluxSigma, as the flux measurement

Table 1: `DIASource` Table

Name	Type	Unit	Description
psFlux	float	nmgy[26]	Calibrated flux for point source model. Note this actually measures the flux *difference* between the template and the visit image.
psFluxSigma	float	nmgy	Estimated uncertainty of `psFlux`.
psLnL	float		Natural *log* likelihood of the observed data given the point source model.
trailFlux	float	nmgy	Calibrated flux for a trailed source model[27,28]. Note this actually measures the flux *difference* between the template and the visit image.
trailLength	float	arcsec	Maximum likelihood fit of trail length[29,30].

Continued on next page

algorithm may be more accurate than the detection algorithm.

[26] A "maggie", as introduced by SDSS, is a linear measure of flux; one maggie has an AB magnitude of 0. "nmgy" is short for a nanomaggie. Flux of 0.063 nmgy corresponds to a 24.5$^{\text{th}}$ magnitude star. See §2.3.5 for details.

[27] A *Trailed Source Model* attempts to fit a (PSF-convolved) model of a point source that was trailed by a certain amount in some direction (taking into account the two-snap nature of the visit, which may lead to a dip in flux around the mid-point of the trail). Roughly, it's a fit to a PSF-convolved line. The primary use case is to characterize fast-moving Solar System objects.

[28] This model does not fit for the *direction* of motion; to recover it, we would need to fit the model to separately to individual snaps of a visit. This adds to system complexity, and is not clearly justified by increased MOPS performance given the added information.

[29] Note that we'll likely measure trailRow and trailCol, and transform to trailLength/trailAngle (or trailRa/trailDec) for storage in the database. A stretch goal is to retain both.

[30] TBD: Do we need a separate trailCentroid? It's unlikely that we do, but one may wish to prove it.

Table 1: `DIASource` Table

Name	Type	Unit	Description
trailAngle	float	degrees	Maximum likelihood fit of the angle between the meridian through the centroid and the trail direction (bearing).
trailLnL	float		Natural *log* likelihood of the observed data given the trailed source model.
trailCov	float[6]	various	Covariance matrix of trailed source model parameters.
fpFlux	float	nmgy	Calibrated flux for point source model measured on the visit image centered at the centroid measured on the difference image (forced photometry flux)
fpFluxSigma	float	nmgy	Estimated uncertainty of `fpFlux`.
fpSky	float	nmgy/asec^2	Estimated sky background at the position (centroid) of the object.
fpSkySigma	float	nmgy/asec^2	Estimated uncertainty of `fpSky`.
E1	float		Adaptive e_1 shape measure of the source as measured on the difference image[31].
E2	float		Adaptive e_2 shape measure.
E1E2cov	float[3]		`E1`, `E2` covariance matrix.
mSum	float		Sum of second adaptive moments.
mSumSigma	float		Uncertainty in `mSum`

Continued on next page

[31]See Bernstein & Jarvis (2002) for detailed discussion of all adaptive-moment related quantities, or `http://ls.st/5f4` for a brief summary.

Table 1: `DIASource` Table

Name	Type	Unit	Description
extendedness	float		A measure of extendedness, computed using a combination of available moments and model fluxes or from a likelihood ratio of point/trailed source models (exact algorithm TBD). $extendedness = 1$ implies a high degree of confidence that the source is extended. $extendedness = 0$ implies a high degree of confidence that the source is point-like.
flags	bit[64]	bit	Flags

2.3.2 `DIAObject` Table

Table 2: `DIAObject` Table

Name	Type	Unit	Description
diaObjectId	uint64		Unique identifier
radec	double[2]	degrees	(α, δ) position of the object at time `radecTai`
radecCov	float[3]	various	`radec` covariance matrix
radecTai	double	time	Time at which the object was at a position `radec`.
pm	float[2]	mas/yr	Proper motion vector[32]
parallax	float	mas	Parallax
pmParallaxCov	float[6]	various	Proper motion - parallax covariances.

Continued on next page

[32]High proper-motion or parallax objects will appear as "dipoles" in difference images. Great care will have to be taken not to misidentify these as subtraction artifacts.

Table 2: `DIAObject` Table

Name	Type	Unit	Description
psFlux	float[ugrizy]	nmgy	Weighted mean point-source model magnitude.
psFluxErr	float[ugrizy]	nmgy	Standard error of `psFlux`
psFluxSigma	float[ugrizy]	nmgy	Standard deviation of the distribution of `psFlux`.
fpFlux	float[ugruzy]	nmgy	Weighted mean forced photometry flux.
fpFluxErr	float[ugrizy]	nmgy	Standard error of `fpFlux`
fpFluxSigma	float[ugrizy]	nmgy	Standard deviation of the distribution of `fpFlux`.
lsPeriod	float[ugrizy]	day	Period (the coordinate of the highest peak in Lomb-Scargle periodogram)
lsSigma	float[ugrizy]	day	Width of the peak at `lsPeriod`.
lsPower	float[ugrizy]		Power associated with `lsPeriod` peak.
lcChar	float[$6 \times M$]		Light-curve characterization summary statistics (eg., 2nd moments, etc.). The exact contents, and an appropriate value of M, are to be determined in consultation with time-domain experts.
nearbyObj	uint64[3]		Closest `Objects` in Level 2 database.
nearbyObjDist	float[3]	arcsec	Distances to `nearbyObj`.
flags	bit[64]	bit	Flags

2.3.3 `SSObject` Table

Table 3: `SSObject` Table

Name	Type	Unit	Description
ssObjectId	uint64		Unique identifier
oe	double[7]	various	Osculating orbital elements at epoch (q, e, i, Ω, ω, M_0, epoch)
oeCov	double[21]	various	Covariance matrix for `oe`
arc	float	days	Arc of observation.
orbFitLnL	float		Natural log of the likelihood of the orbital elements fit.
nOrbFit	int16		Number of observations used in the fit.
MOID	float[2]	AU	Minimum orbit intersection distances[33]
moidLon	double[2]	degrees	MOID longitudes.
H	float[6]	mag	Mean absolute magnitude, per band.
G	float[6]	mag	Fitted slope parameter, per band[34]
hErr	float[6]	mag	Uncertainty in estimate of H
gErr	float[6]	mag	Uncertainty in estimate of G
flags	bit[64]	bit	Flags

The LSST database will provide functions to compute the phase (Sun-Asteroid-Earth) angle α for every observation, as well as the reduced ($H(\alpha)$) and absolute (H) asteroid magnitudes.

[33] `http://www2.lowell.edu/users/elgb/moid.html`

[34] The slope parameter for the large majority of asteroids will not be well constrained until later in the survey. We may decide not to fit for it at all over the first few DRs, and add it later in Operations. Alternatively, we may fit it with a strong prior.

2.3.4 Estimator and Naming Conventions

We employ a convention where estimates of standard errors have the suffix `Err`, while the estimates of inherent widths of distribution (or functions in general) have the suffix `Sigma`[35]. The latter are defined as the square roots of the second moment about the quoted value of the quantity at hand.

Unless noted otherwise, maximum likelihood values are be quoted for all fitted parameters (measurements). Together with covariances, these let the end-user apply whatever prior they deem appropriate when computing posteriors[36].

For fluxes, we recognize that a substantial fraction of astronomers will just want the posteriors marginalized over all other parameters, trusting the LSST experts to select an appropriate prior[37]. For example, this is nearly always the case when constructing color-color or color-magnitude diagrams. We will support these use cases by providing additional pre-computed columns, taking care to name them appropriately so as to minimize accidental incorrect usage. For example, a column named `gFlux` may be the expectation value of the g-band flux, while `gFluxML` may be the maximum likelihood value.

2.3.5 Fluxes and Magnitudes

Because flux measurements on difference images are performed against a template, the measured flux of a source on the difference image can be negative. The flux can also go negative for faint sources in the presence of noise. Negative fluxes cannot be stored as (Pogson) magnitudes; log of a negative number is undefined. We've therefore decided to store fluxes rather than magnitudes, in database tables[38].

We quote fluxes in units of "maggie". A maggie, as introduced by SDSS, is a linear measure of flux. An object with flux of one maggie (integrated over the bandpass) has an AB magnitude of zero:

$$m_{AB} = -2.5 \log_{10}(f/\text{maggie}) \tag{1}$$

[35]Given N measurements, standard errors scale as $N^{-1/2}$, while widths remain constant.

[36]There's a tacit assumption that a Gaussian is a reasonably good description of the likelihood surface around the ML peak.

[37]It's likely that most cases will require just the expectation value alone.

[38]This is a good idea in general. Eg. given multi-epoch observations, one should always be averaging fluxes, rather than magnitudes.

We chose to use maggies (as opposed to Jansky) to allow the user to differentiate between two separate sources of calibration error: the error in relative (internal) calibration of the survey, and the error in absolute calibration that depends on the knowledge of absolute flux of photometric standards.

We realize that the large majority of users will want to work with magnitudes. For convenience, we plan to provide columns with (Pogson) magnitudes[39], where values with negative flux will evaluate to `NULL`. Similarly, we will provide columns with flux expressed in Jy (and its error estimates).

2.3.6 Precovery Measurements

When a new `DIASource` is detected, it's useful to perform (PSF) forced photometry at the location of the new source on images taken prior to discovery. These are colloquially know as *precovery measurements*[40]. Performing precovery in real time over all previously acquired visits is too I/O intensive to be feasible. We therefore plan the following:

1. For all newly discovered objects, perform precovery PSF forced photometry on visits taken over the previous 30 days[41].

2. Make available a "precovery service" to request precovery for a limited number of `DIASources` across all previous visits, and make it available within 24 hours of the request. Web interface and machine-accessible APIs will be provided.

The former should satisfy the most common use cases (eg., SNe), while the latter will provide an opportunity for more extensive timely precovery of targets of special interest.

2.3.7 Annual Reprocessings

In what we've described so far, the Level 1 database is continually being added to as new images are taken and `DIASources` identified. Every time a new `DIASource` is associated to an existing `DIAObject`, the `DIAObject` record

[39] These will most likely be implemented as "virtual" or "computed" columns

[40] When Solar System objects are concerned, precovery has a slightly different meaning: predicting the position of a newly discovered `SSObject` on previously acquired visits, and associating with it `DIASources` consistent with its predicted position.

[41] We will be maintaining a cache of 30 days of processed images to support this feature.

is updated to incorporate new information brought in by the `DIASource`. Once discovered and measured, the `DIASources` would never be re-discovered and re-measured at the pixel level.

This is not optimal. Newer versions of LSST pipelines will improve detection and measurements on older data. Also, PSF forced photometry should be performed on the position of the `DIAObject` on all pre-discovery images. This argues for periodic *reprocessing* of the Level 1 data set.

We plan to reprocess all image differencing-derived data (the Level 1 database), at the same time as we perform the annual Level 2 data release productions. This will include all images taken since the start of observation, to the time when the DR production begins. The images will be reprocessed using a single version of the image differencing and measurement software, resulting in a consistent data set.

As reprocessing is expected to take approximately ~ 9 months, more imaging will be acquired in the meantime. These data will be reprocessed as well, and added to the new Level 1 database generated by the data release processing. The reprocessed database will thus "catch up" with the Level 1 database currently in use, possibly in a few increments. Once it does, the existing Level 1 database will be replaced with the new one, and all future alerts will refer to the reprocessed Level 1 database. Alerts for new sources "discovered" during data release processing and/or the catch-up process will *not* be issued.

Note that Level 1 database reprocessing and switch will have *significant* side-effects on downstream users. For example, all `DIASource` and `DIAObject` IDs will change in general. Some `DIASources` and `DIAObjects` will disappear (eg., if they're image subtraction artifacts that the improved software was now able to recognize as such). New ones may appear. The `DIASource`/`DIAObject`/`Objects` associations will change as well.

While the annual database switches will undoubtedly cause technical inconvenience (eg., a `DIASource` detected at some position and associated to one `DIAObject` ID on day $T - 1$, will now be associated to a different `DIAObject` ID on day $T + 0$), the resulting database will be a more accurate description of the astrophysics that the survey is seeing (eg., the association on day $T + 0$ is the correct one; the associations on $T - 1$ and previous days were actually made to an artifact that skewed the `DIAObject` summary of measurements).

To ease the transition, third parties (event brokers) may choose to provide positional-crossmatching to older versions of the Level 1 database. A set of best practices will be developed to minimize the disruptions caused by the switches (eg., when writing event-broker queries, filter on position, not on `DIAObject` ID, if possible, etc.). A Level 1 database distribution service, allowing for bulk downloads of the reprocessed Level 1 database, will need to be established to support the brokers who will use it locally to perform more advanced brokering[42].

Older versions of the Level 1 database will be archived following the same rules as for the Level 2 databases. DR1, the most recent DR, and the penultimate data release will be kept on disk and loaded into the database. Others will be archived to tape and available as bulk downloads.

2.3.8 Repeatability of Queries

We require that queries executed at a known point in time against some version of the Level 1 database be repeatable at a later date. The exact implementation of this requirement is under consideration by the DM database team.

One possibility may be to make the key tables (nearly) append-only, with each row having two timestamps – createdTai and deletedTai, so that queries may be limited through a `WHERE` clause:

```
SELECT * FROM DIASource WHERE 'YYYY-MM-DD-HH-mm-SS' BETWEEN
createdTAI and deletedTAI
```

or, more generally:

```
SELECT * FROM DIASource WHERE ''data is valid as of YYYY-MM-DD"
```

A perhaps less error-prone alternative, if technically feasible, may be to provide multiple virtual databases that the user would access as:

```
CONNECT lsst-dr5-yyyy-mm-dd
SELECT * FROM DIASource
```

The latter method would probably be limited to nightly granularity, unless there's a mechanism to create virtual databases/views on-demand.

[42] A "bulk-download" database distribution service will be provided for the Level 2 databases as well, to enable end-users to establish and run local mirrors (partial or full).

2.3.9 Uniqueness of IDs across database versions

To reduce the likelihood for confusion, all `Source`, `Object`, `DIASource`, and `DIAObject` IDs shall be unique across database versions. For example, DR4 and DR5 reprocessings will share no identical IDs.

Note, however, that exposure and visit IDs will remain the same across releases.

2.4 Level 1 Image Products

2.4.1 Visit Images

Raw and processed visit images will be made available for download no later than 300 seconds from the end of visit acquisition.

The images will remain accessible with low-latency (seconds from request to start of download) for at least 30 days, with slower access afterwards (minutes to hours).

2.4.2 Difference Images

Complete difference images will be made available for download no later than 300 seconds from the end of visit acquisition.

The images will remain accessible with low-latency (seconds from request to start of download) for at least 30 days, with slower access afterwards (minutes to hours).

2.4.3 Image Differencing Templates

Templates for difference image analysis will be created by coadding 6-months to a year long groups of visits. The coaddition process will take care to remove any transients or fast moving objects (eg., asteroids) from the templates.

The input images may be further grouped by airmass and/or seeing[43]. Therefore, at DR11, we will be creating 11 groups templates: two for the first year of the survey (DR1 and DR2), and then one using imaging from each subsequent year.

Difference image analysis will use the appropriate template given the time of observation, airmass, and seeing.

[43]The number and optimal parameters for airmass/seeing bins will be determined in Commissioning.

2.5 Alerts to DIASources

2.5.1 Information Contained in Each Alert

For each detected DIASource, LSST will emit an "Event Alert" within 60 seconds of the end of visit. These alerts will be issued in VOEvent format[44], and should be readable by VOEvent-compliant clients.

Each alert (a VOEvent packet) will at least include the following:

- Level 1 database id (example: DR5-Level1)
- alertTimestamp (A timestamp that can be used to execute a query against the Level 1 database as it existed when this alert was issued)
- Science Data:
  - The DIASource record that triggered the alert
  - The entire DIAObject (or SSObject) record
  - All previous DIASource records
- 30×30 pixel cut-out of the difference image (10 bytes/pixel, FITS MEF)
- 30×30 pixel cut-out of the template image (10 bytes/pixel, FITS MEF)

2.5.2 Receiving and Filtering the Alerts

Alerts will be transmitted in VOEvent format, using standard IVOA protocols (eg., VOEvent Transport Protocol; VTP). As a very high rate of alerts is expected, approaching ~ 2 million per night, we plan for public VOEvent Event Brokers[45] to be the primary end-points of LSST's VTP streams. End-users will use these brokers to classify and filter events on the stream for

[44] Or some other format that is broadly accepted and used by the community at the start of LSST commissioning.

[45] These brokers are envisioned to be operated as a public service by third parties who will have signed MOUs with LSST. An example may be the VAO or its successor.

those fitting their science goals. End-users will *not* be able to subscribe to full, unfiltered, alert streams coming directly from LSST[46].

For the end-users, LSST will provide a basic, limited capacity, alert filtering service. This service will run at the LSST archive center (at NCSA). It will let astronomers create simple filters that limit what alerts are ultimately forwarded to them[47]. These *user defined filters* will be possible to specify using an SQL-like declarative language, or short snippets of (likely Python) code. For example, here's what a filter may look like:

```
# Keep only never-before-seen events within two
# effective radii of a galaxy. This is for illustration
# only; the exact methods/members/APIs may change.

def filter(alert):
    if len(alert.sources) > 1:
        return False
    nn = alert.diaobject.nearest_neighbors[0]
    if not nn.flags.GALAXY:
        return False
    return nn.dist < 2. * nn.Re
```

We emphasize that this LSST-provided capability will be limited, and is *not* intended to satisfy the wide variety of use cases that a full-fledged public Event Broker could. For example, we do not plan to provide any classification (eg., "is the light curve consistent with an RR Lyra?", or "a Type Ia SN?"). No information beyond what is contained in the `VOEvent` packet will be available to user-defined filters (eg., cross-matches with other catalogs). The complexity and run time of user defined filters will be limited by available resources. Execution latency will not be guaranteed. The number of `VOEvents` transmitted to each user per user will be limited as well (eg., at least up to ~ 20 per visit per user, dynamically throttled depending on load). Finally, the total number of simultaneous subscribers is likely to be

[46]This is due to finite network bandwidth available: for example, a 100 end-users subscribing to a ~ 100 Mbps stream (the peak full stream data rate at end of the first year of operations) would require 10Gbps WAN connection from the archive center, just to serve the alerts.

[47]More specifically, to their VTP clients. Typically, a user will use the Science User Interface (the web portal to LSST archive center) to set up the filters, and use their VTP client to receive the filtered `VOEvent` stream.

limited – in case of overwhelming interest, a TAC-like proposal process may be instituted.

2.6 Open Issues

What follows is a (non-exhaustive) list of issues, technical and scientific, that are still being discussed and where changes are likely. Input on any of these will be appreciated. These need to be resolved before this document is baselined.

- *What light-curve metric should we compute and provide with alerts?* We strive to compute general purpose metrics which will facilitate classification. We have not baselined any yet.

- *Should we measure on individual snaps (or their difference)?* Is there a demonstrable science case requiring immediate followup that would be triggered by the flux change over a ∼15 second period? Is it technically feasible?

- *Should we choose* `nearbyObj`*s differently?* One proposal is to find the brightest `Object` within XX arcsec (with $XX \sim$ 10arcsec), and the total number of `Objects` within XX arcsec.

- *Should the postage stamps provided with the alerts be binned, and by what factor?*

- *When should we (if ever) stop performing forced photometry on positions of* `DIAObjects`*?* Depending on the rate of false positives, unidentified artifacts, or unrecognized Solar System objects, the number of forced measurements may dramatically grow over time.

- *Can we, should we, and how will we measure proper motions on difference images?* This is a non-trivial task (need to distinguish between dipoles that are artifacts, and those due to proper motions), without a clear science driver (since high proper motion stars will be discoverable using Level 2 catalogs).

- *Is Level 1 database required to be relational?*. A no-SQL solution may be more appropriate given the followup-driven use cases. Even if it is relational, the Level 1 database will *not* be sized or architected to perform well on large or complex queries (eg. complex joins, full table scans, etc.).

- *Can users query the Level 1 database for all* ***`DIASources`*** *next to an* ***`Object`****?* Is this technically feasible?

3 Level 2 Data Products

3.1 Overview

Level 2 data products result from direct image[48] analysis. They're designed to enable *static sky* science (eg., studies of galaxy evolution, or weak lensing), and time-domain science that is not time sensitive (eg. statistical investigations of variability). They include image products (reduced single-epoch exposures, called *calibrated exposures*, and coadds), and catalog products (tables of objects, sources, their measured properties, and related metadata).

Similarly to Level 1 catalogs of `DIAObjects` and `DIASources`, `Objects` in the Level 2 catalog represent the astrophysical phenomena (stars, galaxies, quasars, etc.), while `Sources` represent their single-epoch observations. `Sources` are independently detected and measured in single epoch exposures and recorded in the `Source` table.

The master list of `Objects` is generated by associating and deblending the list of single-epoch source detections and the lists of sources detected on special-purpose built *deep coadds*. The coadds used for `Object` detection are designed to be deeper and have better effective seeing than the median visit. At least one coadd per band will be built, a multi-color coadd, and possibly a series of shorter period (eg. yearly) coadds[49]. The flux limit in deep coadds will be significantly fainter than on single visits, and detected sources will be easier to deblend.

The deblender will be run simultaneously on the catalog of peaks[50] detected in the deep coadds, the `DIAObject` catalog from the Level 1 database, and one or more external catalogs. It will use the knowledge of peak positions, bands, time, time variability (from Level 1 and the single-epoch `Source` detections), inferred motion, Galactic longitude and latitude, and other available information to produce a master list of deblended `Objects`. Metadata

[48] As opposed to *difference image*, in Level 1.

[49] The short-period coadds are necessary to avoid missing faint objects showing long-term variability. The short-period coadds will not be preserved. We will provide a facility to regenerate small subsections of the coadds, up to few $\deg^2$, as Level 3 tasks.

[50] The source detection algorithm we plan to employ finds regions of connected pixels above the nominal S/N threshold in the *PSF-likelihood image* of the visit (or coadd). These regions are called *footprints*. Each footprint may have one or more *peaks*, and it is these peaks that the deblender will use to infer the number and positions of objects blended in each footprint.

on why and how a particular `Object` was deblended will be kept.

The properties of `Objects`, including their exact positions, motions, parallaxes, and shapes, will be characterized by MultiFit-type algorithms[51].

Finally, to enable studies of variability, the fluxes of all `Objects` will be measured on individual epochs while keeping their shape parameters and deblending resolutions constant. This process is known as *forced photometry*, and the flux measurements will be stored in the `ForcedSource` table.

3.2 Level 2 Data Processing

Figure 1 presents a high-level view of the Level 2 data processing workflow[52]. Logically[53], the processing begins with single-frame (visit) image reduction and source measurement, followed by global astrometric and photometric calibration, coadd creation, detection on coadds, association and deblending, object characterization, and forced photometry measurements.

The following is a high-level description of steps which will occur during regular Level 2 data processing:

1. *Single Frame Processing*: Raw exposures are reduced to *calibrated visit exposures*, and `Sources` are independently detected, deblended, and measured on all visits. Their measurements (instrumental fluxes and shapes) are stored in the `Source` table.

2. *Relative calibration*: The survey is internally calibrated, both photometrically and astrometrically. Relative zero point and astrometric corrections are computed for every visit. Sufficient data is kept to reconstruct the normalized system response function $\phi_b(\lambda)$ (see Eq. 5, SRD) at every position in the focal plane at the time of each visit as required by § 3.3.4 of the SRD.

3. *Coadd creation*: Deep per-band coadds are created in *ugrizy* bands, as well as deeper, multi-color, coadds[54]. They will be optimized for a

[51] "MultiFit algorithms" are those that fit a PSF-convolved model to all multi-epoch observations of an object. This is in contrast to measurement techniques where multi-epoch images are coadded first, and the properties are measured from the coadded pixels.

[52] Some LSST documents refer to *Data Release Processing*, which includes both Level 1 reprocessing (see § 2.3.7), and the Level 2 processing described here

[53] The actual implementation may parallelize these steps as much as possible.

[54] We'll denote the "band" of the multi-color coadd as 'M'.

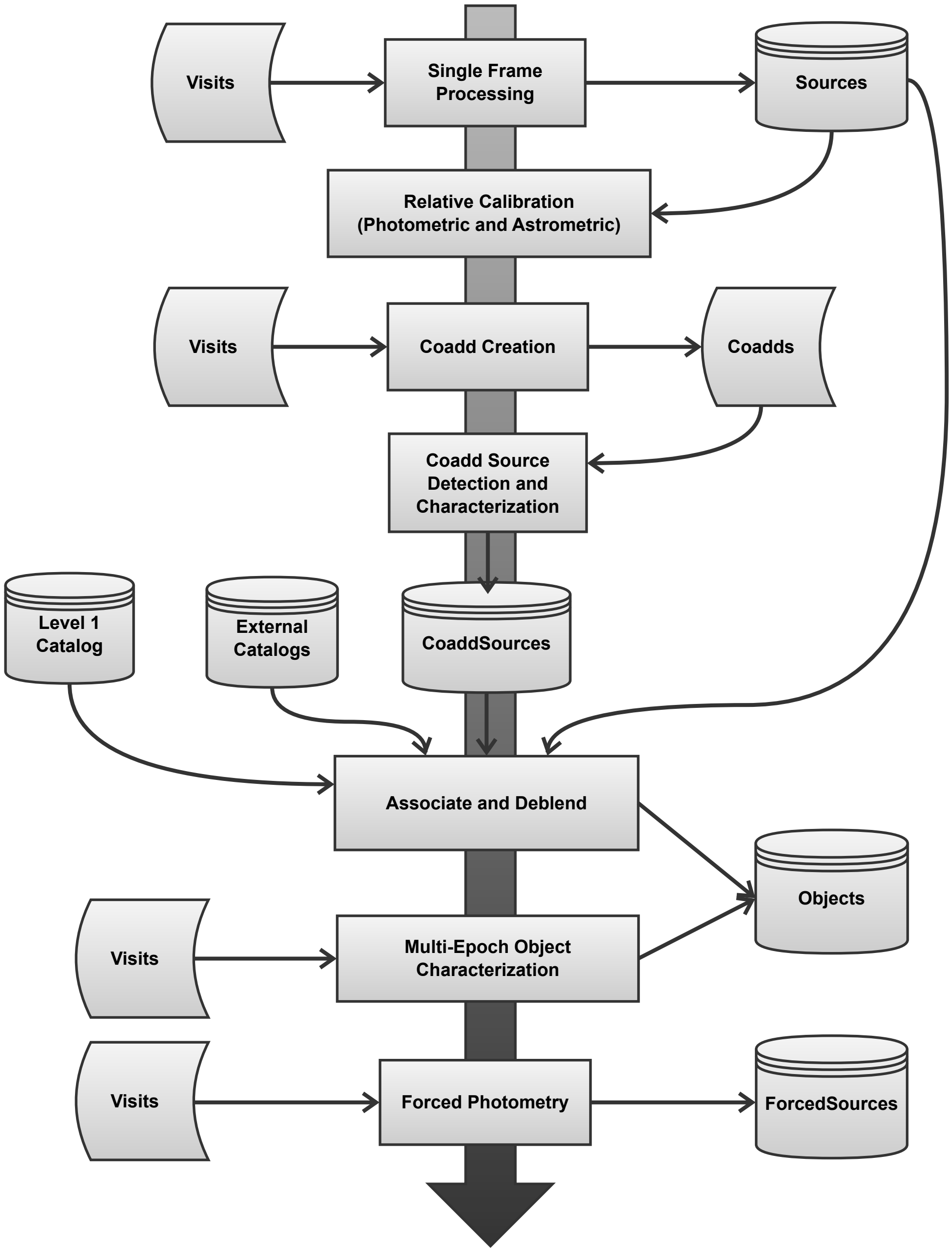

Figure 1: Level 2 Data Processing Overview

reasonable combination of depth (i.e., no PSF matching) and resolution (i.e., only above-average seeing visits may be used to construct them). Transient sources (including Solar System objects, explosive transients, etc), will be rejected from the coadds. See § 3.4.3 for details.

4. *Coadd source detection and characterization.* Sources will be detected on all coadds generated in the previous step. The source detection algorithm will detect regions of connected pixels, known as *footprints*, above the nominal S/N threshold in the *PSF-likelihood image* of the visit. Each footprint may have one or more *peaks*, and the collection of these peaks (and their membership in the footprints) are the output of this stage. This information will be stored in a catalog of `CoaddSources`[55].

5. *Association and deblending.* The next stage in the pipeline, which we will for simplicity just call *the deblender*, synthesizes a list of unique objects. In doing so it will consider the catalog of `Sources`, the the catalog of `CoaddSources`, `DIAObjects` and `DIASources` detected on difference images, and objects from external catalogs.

   The deblender will make use of all information available at this stage, including the knowledge of peak positions, bands, time, time variability (from Level 1), Galactic longitude and latitude, etc. The output of this stage is a list of (uncharacterized) `Objects`[56].

6. *Multi-epoch object characterization.* A set of predefined model fits and measurements will be performed on each of the `Objects` identified in the previous step, taking all available multi-epoch data into account. Model fits will be performed using *MultiFit*-type algorithms. Rather than coadding a set of images and measuring object characteristics on the coadd, MultiFit simultaneously fits PSF-convolved models to the objects multiple observations. This reduces systematic errors, improves the overall S/N, and allows for fitting of time-dependent quantities degenerate with shape on the coadds (for example, the proper motion).

[55] The exact contents of this catalog will be implementation specific and will not be described here.

[56] Depending on the exact implementation of the deblender, this stage may also attach significant metadata (eg, deblended footprints and pixel-weight maps) to each deblended `Object` record.

The models we plan to fit will *not* allow for flux variability (see the next item).

7. *Forced Photometry.* Source fluxes are measured at every visit, with the position, motion, shape, and the deblending parameters characterized in the previous step kept fixed. This process of *forced photometry*, results in characterization of the light-curve for each object in the survey. The fluxes will be stored in the `ForcedSource` table.

3.2.1 Object Characterization Measures

Properties of detected objects will be measured as a part of the object characterization step described in the previous section and stored in the `Object` table. These measurements are designed to enable LSST "static sky" science. This section discusses at a high level which properties will be measured and how those measurements will be performed. For a detailed list of quantities being fit/measured, see the table in § 3.3.1.

All measurements discussed in this section deal with properties of *objects*, and will be performed on multi-epoch coadds, or by simultaneously fitting to all epochs. Measurements of sources in individual visits, independent of all others, are described in § 3.2.2.

To enable science cases depending on observations of non-variable objects in the LSST-observed sky, we plan to measure the following using the MultiFit approach:

- *Point source model fit.* The observed object is modeled as a point source with finite proper motion and parallax and constant flux (a-priori different in each band). This model is a good description for stars and other unresolved sources. Its 11 parameters will be simultaneously constrained using information from all available observations in all bands[57].

- *Bulge-disk model fit.* The object is modeled as a sum of a de Vaucouleurs (Sersic $n = 4$) and an exponential (Sersic $n = 1$) component. This model is intended to be a reasonable description of galaxies[58]. The

[57] The fitting procedure will account for differential chromatic refraction.

[58] We may reconsider this choice if a better suited parametrization is discovered while LSST is in Construction.

object is assumed not to move[59]. The components share the same ellipticity and center. One effective radius is fit for each component (that is, the radius is *not* a function of band). The central surface brightness is allowed to vary from band to band. There are a total of 18 free parameters, which will be simultaneously constrained using information from all available epochs and bands. Where there's insufficient data to constrain the likelihood (eg., small, poorly resolved, galaxies, or very few epochs), priors will be adopted to limit the range of its sampling.

In addition to the maximum likelihood values of fitted parameters and their covariances, a number (currently planned to be ~ 200, on average) of independent samples from the likelihood function will be provided. These will enable use-cases sensitive to departures from the Gaussian approximation.

- *Standard colors.* Colors of the object in "standard seeing" (for example, the third quartile expected survey seeing in the i band, ~ 0.8") will be measured. These colors are guaranteed to be seeing-insensitive, suitable for estimation of photometric redshifts[60].

- *Centroids.* Centroids will be computed independently in each band using an algorithm similar to that employed by SDSS. Information from all epochs will be used to derive the estimate. These centroids will be used for adaptive moment, Petrosian, Kron, standard color, and aperture measurements.

- *Adaptive moments.* Adaptive moments will be computed using information from all epochs, independently for each band. The moments of the PSF realized at the position of the object will be provided as well.

- *Petrosian and Kron fluxes.* Petrosian and Kron radii and fluxes will be measured in standard seeing using self-similar elliptical apertures computed from adaptive moments. The apertures will be PSF-corrected and *homogenized*, convolved to a canonical circular PSF[61]. The radii

[59]I.e., have zero proper motion.

[60]The problem of optimal determination of photometric redshift is the subject of intense research in the community. The approach we're taking here is conservative, following contemporary practices. As new insights develop, we will revisit the issue.

[61]This is an attempt to derive a definition of elliptical apertures that does not depend on seeing. For example, for a large galaxy, the correction to standard seeing will introduce

will be computed independently for each band. Fluxes will be computed in each band, by integrating the light within some multiple of *the radius measured in the canonical band*[62] (most likely the i band). Radii enclosing 50% and 90% of light will be provided.

- *Aperture surface brightness.* Aperture surface brightness will be computed in a variable number[63] of concentric, logarithmically spaced, PSF-homogenized, elliptical apertures, in standard seeing.

3.2.2 Source Characterization

Sources will be detected on individual visits as well as the coadds. Sources detected on coadds will primarily serve as inputs to the construction of the master object list as described in § 3.2, and are not intended to directly support any of LSST science cases.

The following source properties are planned to be measured:

- *Static point source model fit.* The source is modeled as a static point source. There are a total of 3 free parameters (α, δ, flux). This model is a good description of stars and other unresolved sources.

- *Centroids.* Centroids will be computed using an algorithm similar to that employed by SDSS. These centroids will be used for adaptive moment and aperture magnitude measurements.

- *Adaptive moments.* Adaptive moments will be computed. The moments of the PSF realized at the position of the object will be provided as well.

little change to measured ellipticity. Corrected apertures for small galaxies will tend to be circular (due to smearing by the PSF). In the intermediate regime, this method results in derived apertures that are relatively seeing-independent. Note that this is only the case for *apertures*; the measured flux will still be seeing dependent and it is up to the user to take this into account.

[62]The shape of the aperture in all bands will be set by the profile of the galaxy in the canonical band alone. This procedure ensures that the color measured by comparing the flux in different bands is measured through a consistent aperture. See `http://www.sdss.org/dr7/algorithms/photometry.html` for details.

[63]The number will depend on the size of the source.

- *Aperture surface brightness.* Aperture surface brightness will be computed in a variable number[64] of concentric, logarithmically spaced, PSF-homogenized, elliptical apertures.

Note that we do *not* plan to fit extended source Bulge+Disk models to individual `Sources`, nor measure per-visit Petrosian or Kron fluxes. These are object properties that are not expected to vary in time[65], and will be better characterized by MultiFit (the `Object` table).

3.2.3 Forced Photometry

Forced Photometry is the measurement of flux in individual visits, given a fixed position, shape, and the deblending parameters of an object. It enables the study of time variability of an object's flux, irrespective of whether the flux in any given individual visit is above or below the detection threshold.

Forced photometry will be performed on all visits, for all `Objects`. The measured fluxes will be stored in the `ForcedSources` table. Due to space constraints, we only plan to measure the PSF flux.

3.2.4 Crowded Field Photometry

A fraction of LSST imaging will cover areas of high object (mostly stellar) density. These include the Galactic plane, the Large and Small Magellanic Clouds, and a number of globular clusters (among others).

LSST does *not* plan to build or deploy algorithms specifically optimized for crowded field photometry. **Processing these areas with advanced crowded field photometry codes is left to the users as a Level 3 task.**

Nevertheless, high-density areas *will* be processed on a *best effort* basis. LSST image processing and measurement software, primarily designed to operate in non-crowded regions, will be minimally extended to perform in areas of crowding. The current LSST applications development plan envisions making the deblender aware of Galactic longitude and latitude, and

[64] The number will depend on the size of the source.

[65] Objects that do change shape with time would, obviously, be of particular interest. Aperture fluxes provided in the `Source` table should suffice to detect these. Further per-visit shape characterization can be performed as a Level 3 task.

permitting it to use that information as a prior when deciding how to deblent objects[66]. While not guaranteed to reach the accuracy or completeness of purpose-built crowded field photometry codes, we expect this approach will yield reasonable results even in areas of high crowding.

3.3 The Level 2 Catalogs

This section presents the contents of key Level 2 catalog tables. As was the case for Level 1 (see § 2.3), here we present the *conceptual schemas* for the most important Level 2 tables.

These convey *what* data will be recorded in each table, rather than the details of *how*. For example, columns whose type is an array (eg., `radec`) may be expanded to one table column per element of the array (eg., `ra`, `decl`) once this schema is translated to SQL. Secondly, the tables to be presented are normalized (i.e., contain no redundant information). For example, since the band of observation can be found by joining a `Source` table to the table with exposure metadata, there's no column for 'band' in the `Source` table. In the as-built database, the views presented to the users will be appropriately denormalized for ease of use.

3.3.1 The `Object` Table

Table 4: Level 2 Catalog `Object` Table

Name	Type	Unit	Description
objectId	uint64		Unique object identifier
psRadecTai	double	time	Point source model: Time at which the object was at position `radec`.
psRadec	double[2]	degrees	Point source model: (α, δ) position of the object at time `radecTai`.
psPm	float[2]	mas/yr	Point source model: Proper motion vector.

Continued on next page

[66]If the deblender knows that virtually all objects are stars (point sources), they can be deblended in a fashon that broadly mimics crowded field photometry codes (eg. `DAOPHOT`).

Table 4: Level 2 Catalog `Object` Table

Name	Type	Unit	Description
psParallax	float	mas	Point source model: Parallax.
psFlux	float[ugrizy]	nmgy	Point source model fluxes[67].
psCov	float[66]	various	Point-source model covariance matrix[68].
psLnL	float		Natural *log* likelihood of the observed data given the point source model.
bdRadec	double[2]	degrees	B+D model[69]: (α, δ) position of the object at time `radecTai`, in each band.
bdEllip	float[2]		B+D model: Ellipticity (e_1, e_2) of the object.
bdFluxB	float[ugrizy]	nmgy/as^2	B+D model: Total flux of the de Vaucouleurs component.
bdFluxD	float[ugrizy]	nmgy/as^2	B+D model: Total flux of the exponential component.
bdReB	float	arcsec	B+D model: Effective radius of the de Vaucouleurs profile component.
bdReD	float	arcsec	B+D model: Effective radius of the exponential profile component.
bdCov	float[171]	various	B+D model covariance matrix[70].

Continued on next page

[67]Point source model assumes that fluxes are constant in each band. If the object is variable, `psFlux` will effectively be some estimate of the average flux.

[68]Not all elements of the covariance matrix will be stored with same precision. While the variances will be stored as 32 bit floats ($\sim$ seven significant digits), the covariances may be stored to $\sim$ three significant digits ($\sim 1\%$).

[69]Though we refer to this model as “Bulge plus Disk”, we caution the reader that the decomposition, while physically motivated, should not be taken too literally.

[70]See `psCov` for notes on precision of variances/covariances.

Table 4: Level 2 Catalog `Object` Table

Name	Type	Unit	Description
bdLnL	float		Natural *log* likelihood of the observed data given the bulge+disk model.
bdSamples	float[18][200]		Independent samples of bulge+disk likelihood surface. All sampled quantities will be stored with at least $\sim$ 3 significant digits of precision. The number of samples will vary from object to object, depending on how well the object's likelihood function is approximated by a Gaussian.
stdColor	float[5]	mag	Color of the object measured in "standard seeing". While the exact algorithm is yet to be determined, this color is guaranteed to be seeing-independent and suitable for photo-Z determinations.
stdColorSigma	float[5]	mag	Uncertainty of `extColor`.
radec	double[6][2]	arcsec	Position of the object (centroid), computed independently in each band. The centroid will be computed using an algorithm similar to that employed by SDSS.
radecSigma	double[6][2]	arcsec	Uncertainty of `radec`.

Continued on next page

Table 4: Level 2 Catalog `Object` Table

Name	Type	Unit	Description
E1	float[ugrizy]		Adaptive e_1 shape measure. See Bernstein & Jarvis (2002) for detailed discussion of all adaptive-moment related quantities[71].
E2	float[ugrizy]		Adaptive e_2 shape measure.
E1E2cov	float[ugrizy][3]		`E1`, `E2` covariance matrix.
mSum	float[ugrizy]		Sum of second adaptive moments.
mSumSigma	float[ugrizy]		Uncertainty in `mSum`
m4	float[ugrizy]		Fourth order adaptive moment.
petroRad	float[ugrizy]	arcsec	Petrosian radius (computed using elliptical apertures defined by the adaptive moments).
petroRadSigma	float[ugrizy]	arcsec	Uncertainty of `petroRad`
petroBand	int8		The band of the canonical `petroRad`
petroFlux	float[ugrizy]	nmgy	Petrosian flux within a defined multiple of the canonical `petroRad`
petroFluxSigma	float[ugrizy]	nmgy	Uncertainty in `petroFlux`
petroRad50	float[ugrizy]	arcsec	Radius containing 50% of Petrosian flux.
petroRad50Sigma	float[ugrizy]	arcsec	Uncertainty of `petroRad50`.
petroRad90	float[ugrizy]	arcsec	Radius containing 90% of Petrosian flux.
petroRad90Sigma	float[ugrizy]	arcsec	Uncertainty of `petroRad90`.
kronRad	float[ugrizy]	arcsec	Kron radius (computed using elliptical apertures defined by the adaptive moments)

Continued on next page

[71] Or `http://ls.st/5f4` for a brief summary.

Table 4: Level 2 Catalog `Object` Table

Name	Type	Unit	Description
kronRadSigma	float[ugrizy]	arcsec	Uncertainty of `kronRad`
kronBand	int8		The band of the canonical `kronRad`
kronFlux	float[ugrizy]	nmgy	Kron flux within a defined multiple of the canonical `kronRad`
kronFluxSigma	float[ugrizy]	nmgy	Uncertainty in `kronFlux`
kronRad50	float[ugrizy]	arcsec	Radius containing 50% of Kron flux.
kronRad50Sigma	float[ugrizy]	arcsec	Uncertainty of `kronRad50`.
kronRad90	float[ugrizy]	arcsec	Radius containing 90% of Kron flux.
kronRad90Sigma	float[ugrizy]	arcsec	Uncertainty of `kronRad90`.
apN	int8		Number of elliptical annuli (see below).
apMeanSb	float[6][`apN`]	nmgy/asec^2	Mean surface brightness within an annulus[72].
apMeanSbSigma	float[6][`apN`]	nmgy/asec^2	Standard deviation of `apMeanSb`.
flags	bit[128]	bit	Flags

3.3.2 `Source` Table

`Source` measurements are performed independently on individual visits. They're designed to enable astrometric and photometric relative calibration, variability studies of high signal-to-noise objects, and studies of high SNR objects that vary in position and/or shape (eg., comets).

[72] A database function will be provided to compute the area of each annulus, to enable the computation of aperture flux.

Table 5: Level 2 Catalog `Source` Table

Name	Type	Unit	Description
sourceId	uint64		Unique source identifier[73]
ccdVisitId	uint64		ID of CCD and visit where this source was measured
objectId	uint64		ID of the `Object` this source was associated with, if any.
ssObjectId	uint64		ID of the `SSObject` this source has been linked to, if any.
psFlux	float	nmgy	Calibrated point source model flux.
psXY	float[2]	pixels	Point source model: $(column, row)$ position of the object on the CCD.
psCov	float[6]	various	Point-source model covariance matrix[74].
psLnL	float		Natural *log* likelihood of the observed data given the point source model.
psRadec	double[2]	degrees	Point source model: (α, δ) position of the object, transformed from `psXY`
psCov2	float[6]	various	Point-source model covariance matrix for `psRadec` and `psFlux`.
xy	float[2]	arcsec	Position of the object (centroid), computed using an algorithm similar to that used by SDSS.
xyCov	float[3]		Covariance matrix for `xy`.

Continued on next page

[73]It would be optimal if the source ID is globally unique across all releases. Whether that's realized will depend on technological and space constraints.

[74]Not all elements of the covariance matrix will be stored with same precision. While the variances will be stored as 32 bit floats ($\sim$ seven significant digits), the covariances may be stored to $\sim$ three significant digits ($\sim 1\%$).

Table 5: Level 2 Catalog `Source` Table

Name	Type	Unit	Description
radec	double[2]	arcsec	Calibrated (α, δ) of the source, transformed from `xy`.
radecCov	float[3]	arcsec	Covariance matrix for `radec`.
E1	float		Adaptive e_1 shape measure.
E2	float		Adaptive e_2 shape measure.
E1E2cov	float[3]		`E1`, `E2` covariance matrix.
mSum	float		Sum of second adaptive moments.
mSumSigma	float		Uncertainty in `mSum`
m4	float		Fourth order adaptive moment.
apN	int8		Number of elliptical annuli (see below).
apMeanSb	float[`apN`]	nmgy	Mean surface brightness within an annulus.
apMeanSbSigma	float[`apN`]	nmgy	Standard deviation of `apMeanSb`.
flags	bit[128]	bit	Flags

3.3.3 `ForcedSource` Table

Table 6: Level 2 Catalog `ForcedSource` Table

Name	Type	Unit	Description
objectId	uint64		Unique object identifier
ccdVisitId	uint64		ID of CCD and visit where this source was measured
psFlux	float	nmgy	Point source model flux.
psFluxErr	float	nmgy	Point source model flux error, stored to 1% precision.

Continued on next page

Table 6: Level 2 Catalog `ForcedSource` Table

Name	Type	Unit	Description
flags	bit[8]	bit	Flags

3.4 Level 2 Image Products

3.4.1 Visit Images

Raw, including individual snaps, and processed visit images will be made available for download as FITS files. They will be downloadable both through the human-friendly Science User Interface, as well as machine-friendly APIs.

Required calibration data, processing metadata, and all necessary image processing software will be provided to enable the user to generate bitwise identical processed images from raw images[75].

3.4.2 Calibration Frames

All calibration frames (darks, flats, biases, fringe, etc.) will be preserved and made available for download as FITS files.

3.4.3 Coadded Images

In course of Level 2 processing, three classes of coadds will be created:

- A set of *deep coadds*. One deep coadd will be created for each of the *ugrizy* bands, plus a seventh, deeper, multi-color coadd[76]. These coadds will be optimized for a reasonable combination of depth (i.e., employ no PSF matching) and resolution (i.e., visits with significantly degraded seeing may be ommitted). Transient sources (including Solar System objects, explosive transients, etc), will be removed. Care will be taken to preserve the astrophysical backgrounds[77].

[75] Assuming identically performing software and hardware configuration.

[76] In text to follow, we'll denote the "band" of the multi-color coadd as 'M'.

[77] For example, using "background matching" techniques; `http://ls.st/l9u`

These coadds will be kept indefinitely and made available to the users. *Their primary purpose is to enable the use of alternative object characterization algorithms, and studies of diffuse structures.*

- Multiple (ugrizyM) sets of yearly coadds. Each of these sets will be created using only a year's worth of data, otherwise being the same as the deep coadds described above. These are designed to enable detection of long-term variable or moving[78] objects that would be "washed out" (or rejected) in full-depth coadds. **We *do not plan to keep and make these coadds available*.** We will retain and provide sufficient metadata for users to re-create them using Level 3 or other resources.

- One (ugrizyM) set of PSF-matched coadds. These will be used to measure colors of objects at "standard" seeing. ***We do not plan to keep and make these coadds available***. We will retain and provide sufficient metadata for users to re-create them using Level 3 or other resources.

To build the coadds, we plan to subdivide the sky into 12 overlapping[79] *tracts*, spanning approximately 75×72 degrees. The sky will be stereographically projected onto the tracts[80], and be pixelized into (logical) images 2.0 x 1.9 megapixels in size (3.8 terapixels in all). Physically, these large images will be subdivided into smaller (e.g. $2\text{k} \times 2\text{k}$), non-overlapping, *patches*, though that will be transparent to the users. The users will be able to request arbitrarily chosen regions[81] in each tract, and receive them back as a FITS file.

We reiterate that **not all coadds will be kept and served to the public**[82], though sufficient metadata will be provided to users to recreate them on their own. Some coadds may be entirely "virtual": for example, the PSF-matched coadds could be implemented as ad-hoc convolutions of postage stamps when the colors are measured.

[78]For example, nearby high proper motion stars.

[79]We're planning for 3.5 degrees of overlap, roughly accommodating a full LSST focal plane.

[80]See `https://dev.lsstcorp.org/trac/wiki/DM/SAT/SkyMap` for details.

[81]Up to some reasonable upper size limit; i.e., we don't plan to expect to support creation of 3.8 Tpix FITS files.

[82]The coadds are a major cost driver for storage. LSST Data Management system is currently sized to keep and serve seven coadds, *ugrizyM*, over the full footprint of the survey.

We **will** retain smaller sections of all generated coadds, to support quality assessment and targeted science. Retained sections may be positioned to cover areas of the sky of special interest such as overlaps with other surveys, nearby galaxies, large clusters, etc.

3.5 Data Release Availability and Retention Policies

Products for Data Release 1, the most recent data release, and the penultimate data release will be kept on fast storage, with catalogs loaded into the database. Other releases will be archived to mass storage (tape). The users ***will not be able to perform database queries against archived releases***. They will be made available as bulk downloads in some common format (for example, FITS binary tables).

All raw data used to generate any public data product (raw exposures, calibration frames, telemetry, configuration metadata, etc.) will be kept and made available for download.

3.6 Level 2 Open Issues

- *Which coadds to we keep and serve to the public?* Probably non-PSF matched coadds with CoaddPSF (aka. StackFit PSF). PSF-matched coadds are another option and may be easier for the users to work with.

4 Deep Drilling Data Products

Say something about the deep drilling data products. These will involve special nightly stacks, special image differencing procedures, and separate databases. Roughly, since the details of all DDs won't be known for a while (and will change), this section will lay out the constraints on what can and cannot be done for DD processing.

5 Level 3 Data Products and Capabilities

Level 3 capabilities are envisioned to enable science cases that would greatly benefit from co-location of user processing and/or data within the LSST Archive Center. The high-level requirement for Level 3 is established in § 3.5 of the LSST SRD.

Level 3 capabilities include three separate deliverables:

1. Level 3 Data Products and associated storage resources
2. Level 3 processing resources
3. Level 3 programming environment / framework

Many scientists' work may involve using two or all three of them in concert, but they can each be used independently.

5.1 Level 3 Data Products and Associated Storage Resources

These are data products that are generated by users *on any computing resources anywhere* that are then brought to an LSST Data Access Center (DAC) and stored there. The hardware for these capabilities includes the physical storage and database server resources at the DAC to support them.

For catalog data products, there is an expectation that they can be "federated" with the Level 1 (L1) and Level 2 (L2) catalogs to enable analyses combining them. Essentially this means that either the user-supplied tables include keys from the L1/L2 catalogs that can be used for key-equality-based joins with them (example: a table of custom photometric redshifts for galaxies, with a column of object IDs that can be joined to the L2 Object catalog), or that there are columns that can be used for spatial (or temporal, or analogous) joins against L1/L2 tables. The latter implies that such L3 table's columns must be in the same coordinate system / units as the corresponding L1/L2 columns.

There is no requirement that Level 3 DPs are derived from L1 or L2 other than that they be joinable with them. For instance, a user might have a catalog of radio sources that they might want to bring into federation with the LSST catalogs. That can be thought of as a Level 3 Data Product as long as they have "LSST-ized" it by ensuring compatibility of coordinate, time, etc. measurement systems. Nevertheless, we do expect the majority of L3DPs to be derived from processed LSST data.

There could also be L3 image data products; for example, user-generated coadds with special selection criteria or stacking algorithms.

Any L3DP may have access controls associated with it, restricting read access to just the owner, to a list of people, to a named group of people, or allowing open access.

The storage resources for L3DPs come out of the SRD requirement for 10% of LSST data management capabilities to be devoted to user processing. In general, they are likely to be controlled by some form of a "space allocation committee". Users will probably have some small baseline automatic allocation, beyond which a SAC proposal is needed. The SAC may take into account scientific merit, length of time for which the storage is requested, and openness of the data to others, in setting its priorities.

It is to be decided whether users will be required to provide the code and/or documentation behind their L3DPs, or whether the SAC may include the availability of this supporting information in its prioritization. Obviously if a user intends to make a L3DP public or publish it to a group it will be more important that supporting information be available.

Level 3 data products that are found to be generally useful can be migrated to Level 2. This is a fairly complex process that ultimately involves the project taking responsibility for supporting and running LSST-style code that implements the algorithm necessary to produce the data product (it's not just relabeling an existing L3DP as L2). The project will provide necessary support for such migrations.

5.2 Level 3 Processing Resources

These are project-owned computing resources located at the DACs. They are available for allocation to all users with LSST data rights. They may be used for any computation that involves the LSST data and advances LSST-related science. The distinctive feature of these computing resources is that they are located with excellent I/O connections to the image and catalog datasets at Level 1 and Level 2.

These resorces will, at least, include systems that can carry out traditional batch-style processing, probably similarly configured to those we'll be using for the bulk of LSST production processing. It is to be determined whether any other flavors of hardware would be provided, such as large-memory machines; this is likely to be driven by the project need (or lack thereof) for such resources.

There will be a time allocation committee (TAC) for these resources. Every LSST-data-rights user may get a small default allocation (enough to run test jobs). Substantial allocations will require a scientific justification. Priorities will be based on the science case and, perhaps, also on whether the results of the processing will be released to a larger audience. Requests must specify what special flavors of computing will be needed (e.g., GPUs, large memory, etc.).

A fairly standard job control environment (like Condor), will be available, and users will be permitted to work with it at a low, generic level. They will not be required to use the higher levels of the LSST process control middleware (but they may; see below).

These processing resources can be available for use in any clearly LSST-related scientific work. It is not strictly required that they be used to process LSST data, in this context. For instance, it could be acceptable to run special types of cosmological simulations that are in direct support of an LSST analysis, *if the closeness to the data makes the LSST facility uniquely suitable for such work*. The TAC will take into account in its decisions whether proposed work makes good use of the enhanced I/O bandwidth available to LSST data on these systems.

5.2.1 Level 3 Programming Environment and Framework

As a part of the Level 3 Programming Environment and Framework, the LSST will make available the LSST software stack to users, to aid in the analyses of LSST data.

These analyses can be done on LSST-owned systems (i.e., on the Level 3 processing resources) but also on a variety of supported external systems. We will aim to support common personal Unix flavors (for example, common distributions of Linux and Mac OS X) as well as commonly used cluster and HPC environments. The vision is to enable relatively straightforward use of major national systems such as XSEDE or Open Science Grid, as well as some common commercial cloud environments. The decision of which environments to support will be under configuration control and we will seek advice from the user community. We cannot commit to too many flavors. In-kind contributions of customizations for other environments will be welcome and may provide a role for national labs.

The Level 3 environment is intended, when put to fullest use, to allow users to run their own productions-like runs on bulk image and/or catalog data, with mechanisms for creating and tracking large groups of jobs in a batch system.

The Level 3 environment, in asymptopia, has a great deal in common with the environment that the Project will use to build the Level 2 data releases. It is distinct, however, as supporting it as a tool meant for the end-users imposes additional requirements:

- In order to be successful as a *user* computing environment, it needs to be easy to use. Experience with prior project[83] has shown that if the production computing environment is not envisioned from the start as being shared with users, it will likely evolve into an experts-only tool that is too complicated, or too work-hardened, to serve users well.

- While it is desirable for the production computing to be portable to Grid, cloud, etc. resources, this option might not be exercised in practice and could atrophy. For the user community, it's a far more central capability.

- Not all the capabilities of the LSST production environment need necessarily be exported to the users. LSST-specific capabilities associated with system administration, for instance, are not of interest to end-users.

5.2.2 Migration of Level 3 data products to Level 2

- For the migration to be considered, the creator of the L3DP will need to agree to make their data product public to the entire LSST data-rights community, along with supporting documentation and code. The code at first need not be in the LSST framework or even in an LSST-supported language.

- If the original proponent wrote her/his code in the C++/Python LSST stack environment (the "Level 3 environment"), it will be easier to migrate it to Level 2 (though, obviously, using the same languages/framworks does not guarantee that the code is of production quality).

- If the original code was written in another language or another data processing framework, the project will have the resources to rewrite it to required LSST standards.

- Taking on a new Level 2 DP means that the project is committing to code maintenance, data quality review, space allocation, and continuing production of the new L2DP through DR11.

[83]For example, BaBar.

6 TBD

- Crowded field photometry
- Variability characterization
- Solar System objects in L2
- Solar system object merging with external catalogs
- Mechanisms for updating the document
- Selection functions
- Deblending, children, etc
- Implement K-T's comments
- Understand how to record photometric calibration
- Add required accuracy to all tables