Large Synoptic Survey Telescope
Data Products Definition Document

[To become LSE-163 pending review and CCB approval]

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

for the LSST Project

June 5, 2013

Abstract

This document describes the data products and processing services to be delivered by the Large Synoptic Survey Telescope (LSST).

The LSST will deliver three levels of data products and services. **Level 1** (nightly) data products will include images, difference images, catalogs of sources and objects detected in difference images, and catalogs of Solar System objects. Their primary purpose is to enable rapid follow-up of time-domain events. **Level 2** (annual) data products will include well calibrated single-epoch images, deep coadds, and catalogs of objects, sources, and forced sources, enabling static sky and precision time-domain science. **Level 3** (user-created) data product services will enable science cases that greatly benefit from co-location of user processing and/or data within the LSST Archive Center. LSST will also devote 10% of observing time to programs with special cadence. Their data products will be created using the same software and hardware as Levels 1 and 2. All data products will be made available using user-friendly databases and web services.

*Please direct comments to <mjuric@lsst.org>.

Contents

1 Preface

The purpose of this document is to describe the data products produced by the Large Synoptic Survey Telescope (LSST).

To a future LSST user, it should clarify what catalogs, image data, software, and services they can expect from LSST. To LSST builders, it provides direction on how to flow down the LSST System Requirements Document to system design, sizing, budget and schedule as they pertain to the data products.

Though under strict change control[1], this is a ***living document***. LSST will undergo a period of construction and commissioning lasting no less than seven years, followed by a decade of survey operations. To ensure their continued scientific adequacy, the designs and plans for LSST Data Products will be periodically reviewed and updated.

[1]LSST Docushare handle for this document is `LSE-163`.

2 Introduction

2.1 The Large Synoptic Survey Telescope

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 every night using pairs of 15-second exposures, 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 catalogs including over 38 billion stars and galaxies, that 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[2].

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 temporal changes in the 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 §4.4.3). Measurements will be internally and absolutely calibrated.

The *broad*, *high-level*, requirements for LSST Data Products are given by the *LSST Science Requirements Document*[3] (SRD). This document lays out the *specifics* of what the data products will comprise of, how those data will

[2]Informally known as "Deep Drilling Fields".

[3]LSST Document Handle LPM-17, available at `http://ls.st/srd`

be generated, and when. It serves to inform the flow-down from the LSST SRD through the *LSST System Requirements Document* (the LSR; LSE-29) and the *LSST Observatory System Specifications* (OSS; LSE-30), to the *LSST Data Management System Requirements* (DMSR; LSE-61), the UML model, and the database schema.

2.2 Classes of LSST 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 will be done on nightly or daily basis and result in **Level 1** data products. Level 1 products will include difference images, catalogs of sources detected in difference images (`DIASources`), astrophysical objects[4] these are associated to (`DIAObjects`), and Solar System objects (`SSObjects`[5]). The catalogs will entered into the **Level 1 database** and made available in near real time. Notifications ("alerts") about new `DIASources` will be issued using community-accepted standards within 60 seconds of observation. Level 1 data products are discussed in § 4.

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 results are **Level 2** data products. These products, generated and released annually[6], will include the single-epoch images, deep coadds, catalogs of characterized `Objects` (detected on deep coadds as well as

[4]The LSST has adopted the nomenclature by which single-epoch detections of astrophysical *objects* are called *sources*. The reader is cautioned that this nomenclature is not universal: some surveys call *detections* what LSST calls *sources*, and use the term *sources* for what LSST calls *objects*.

[5]`SSObjects` used to be called call "Moving Objects" in previous versions of the LSST 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.

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

individual visits[7]), `Sources`[8] (detections and measurements on individual visits), and `ForcedSources` (constrained measurement of flux on individual visits). It will also include fully reprocessed Level 1 data products (see §4.3.5). 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. Level 2 data products are discussed in § 5.

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 acquisition. The analysis of science (direct) images is less time sensitive, and will be done as a part of annual data release process.

Recognizing the diversity of astronomical community needs, and the need for specialized processing not part of the automatically generated Level 1 and 2 products, LSST plans to devote 10% of its data management system capabilities to enabling the creation, use, and federation of **Level 3** (user-created) data products. Level 3 capabilities will enable science cases that 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. Their details are discussed in § 6 of this document.

Finally, LSST Survey Specifications (§ 3.4 of LSST SRD) prescribe that 90% of LSST observing time be spent in the so-called "universal cadence" mode of surveying the sky. These observations will result in Level 1 and 2 data products discussed above. The remaining 10% of observing time will be devoted to **special programs**, designed to obtain improved coverage of interesting regions of observational parameter space. Examples include very deep ($r \sim 26$, per exposure) observations, observations with very short revisit times ($\sim$1 minute), and observations of "special" regions such as the Ecliptic, Galactic plane, and the Large and Small Magellanic Clouds. The data products for these programs will be generated using the same processing

[7]The LSST takes two exposures per pointing, nominally 15 seconds in duration each, called *snaps*. For the purpose of data processing, that pair of exposures will typically be coadded and treated as a single exposure, called a *visit*.

[8]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.

software and hardware and possess the general characteristics of Level 1 and 2 data products, but may be performed on a somewhat different cadence. They will be discussed in § 7.

3 General Considerations

Most LSST data products will consist of images and/or catalogs. The catalogs will be stored and offered to the users as *relational databases* which they will be able to query. This approach was shown to work well by prior large surveys, for example the Sloan Digital Sky Survey (SDSS).

Different data products will generally be stored in different databases. For example, Level 1 data products will be stored in a *Level 1 database*. Level 2 "universal cadence" products will be stored in a *Level 2 database* database. The products for special programs may be stored in many different databases, depending on the nature of the program.

Nevertheless, all these databases will follow certain naming and other conventions. We discuss these in the subsections to follow.

3.1 Estimator and Naming Conventions

For all catalogs data, we will 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`[9]. 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 will be quoted for all fitted parameters (measurements). Together with covariances, these let the end-user apply whatever prior they deem appropriate when computing posteriors[10]. Where appropriate, multiple independent samples from the likelihood may be provided to characterize departures from Gaussianity.

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[11]. 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

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

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

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

of the g-band flux, while `gFluxML` may represent the maximum likelihood value.

3.2 Fluxes and Magnitudes

Because flux measurements on difference images (Level 1 data products; § 4) 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 therefore prefer to store fluxes, rather than magnitudes, in database tables[12].

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

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

We chose to use maggies (as opposed to, say, Jansky) to allow the user to differentiate between two distinct sources of photometric 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.

Nevertheless, we acknowledge that the large majority of users will want to work with magnitudes. For convenience, we plan to provide columns with (Pogson) magnitudes[13], where values with negative flux will evaluate to `NULL`. Similarly, we will provide columns with flux expressed in Jy (and its error estimate that includes the relative and absolute calibration error contributions).

3.3 Uniqueness of IDs across database versions

To reduce the likelihood for confusion, all IDs shall be unique across databases and database versions, other than those corresponding to uniquely identifiable entities (i.e., IDs of exposures).

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

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

For example, DR4 and DR5 (or any other) release will share no identical `Object`, `Source`, `DIAObject` or `DIASource` IDs (see § 4 and 5 for the definitions of `Objects`, `DIAObjects`, etc.).

3.4 Repeatability of Queries

We require that queries executed at a known point in time against any LSST-delivered database be repeatable at a later date. This promotes the reproducibility of science derived from LSST data. It is of special importance for Level 1 catalogs (§ 4) that will change on a nightly basis as new time domain data is being processed and added to the catalogs.

The exact implementation of this requirement is left to the LSST Data Management 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 by 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.

4 Level 1 Data Products

4.1 Overview

Level 1 data products are a result of difference image analysis (DIA; §4.2.1). They include the sources detected in difference images (`DIASources`), astrophysical objects that these are associated to (`DIAObjects`), identified Solar System objects[14] (`SSObject`), and related, broadly defined, metadata (including eg., cut-outs[15]).

`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 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). Their shape can be complex (eg., trailed, for a source with proper motion approaching $\sim$ deg/day, or "dipole-like", if an object's observed position exhibits an offset – true or apparent – compared to its position on the template).

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[16], or by assuming it to be distant enough to only exhibit small parallactic and proper motion[17]. 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

[14]The SRD considers Solar System object orbit catalog to be a Level 2 data product (LSST SRD, 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.

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

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

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

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 of each visit, we will issue time domain event alerts for all newly discovered DIASources[18].

4.2 Level 1 Data Processing

4.2.1 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[19], combining of snaps, etc.).

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

3. The flux and shape[20] of the DIASource are measured on the difference image. PSF photometry is performed on the visit image 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 §4.3) is searched for a DIAObject or an SSObject with which to positionally associate the newly discovered DIASource[21]. If no match is found, a new DIAObject is created and the observed DIASource is associated to it.

[18]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.

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

[20]The "shape" in this context consists of weighted 2nd moments, as well as a fit to a trailed source model.

[21]The association algorithm will guarantee that a DIASource is associated with not more than one existing 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.

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 4.2.2).

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[22] is searched for one or more `Objects` positionally close to the `DIAObject`, out to some maximum radius[23]. The IDs of these `Objects` are recorded in the `DIAObject` record and provided in the issued event alert (see below).

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[24], name of the Level 2 database and the IDs of nearby `Objects`, and the associated science content (centroid, fluxes, low-order lightcurve moments, periods, etc.), *including the full light curves*. See Section 4.5 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`[25]. 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.

[22]Level 2 database is a database resulting from annual data release processing. See § 5 for details.

[23]Eg., a few arcseconds.

[24]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).

[25]For the purposes of this document, we're treating the `DIASources` generated by forced photometry or 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.

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[26] 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 hazardous asteroid. The project will define the initial set of criteria by the start of Operations.

4.2.2 Solar System Object Processing

The following will occur during regular Solar System object processing (in daytime[27], after a night of observing):

1. The orbits and physical properties of all `SSObjects` re-observed on the previous night are recomputed. External orbit catalogs (or observations) are used to improve orbit estimates. Updated data are entered to the `SSObjects` table.

2. All `DIASources` detected on the previous night, that have not been matched at a high confidence level 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.[28].

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

[27] Note that there *is no strict bound 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.

[28] 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.

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

4.3 Level 1 Catalogs

The described alert processing design relies on the Level 1 database that contains the objects and sources detected on difference images. At the very least[29], this database will have tables of `DIASources`, `DIAObjects`, and `SSObjects`, populated in the course of difference image and Solar System object processing[30]. As these get updated and added to, their updated contents becomes visible (query-able) immediately[31].

This database is *only loosely coupled to the Level 2 database.* All of the coupling is through 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 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).

[29] It will also contain exposure and visit metadata, MOPS-specific tables, etc. These are either standard/uncontroversial, implementation-dependent, or less directly relevant for science and therefore not discussed in this document.

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

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

- *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 relation (`DIASource-DIAObject-Object`). For ease of use, views or pre-built table with these will be offered to the 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 largely normalized (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 named `band` in the `DIASource` table. In the as-built database, the views presented to the users will be appropriately denormalized for ease of use.

4.3.1 `DIASource` Table

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

Continued on next page

Table 1: `DIASource` Table

Name	Type	Unit	Description
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)[32]
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.[33]
psFlux	float	nmgy[34]	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.

Continued on next page

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

[33]This is not necessarily the same as psFlux/psFluxSigma, as the flux measurement algorithm may be more accurate than the detection algorithm.

[34]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^{th} magnitude star. See §3.2 for details.

Table 1: `DIASource` Table

Name	Type	Unit	Description
trailFlux	float	nmgy	Calibrated flux for a trailed source model[35,36]. Note this actually measures the flux *difference* between the template and the visit image.
trailLength	float	arcsec	Maximum likelihood fit of trail length[37,38].
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)

Continued on next page

[35]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.

[36]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.

[37]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.

[38]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
fpFluxSigma	float	nmgy	Estimated uncertainty of `fpFlux`.
E1	float		Adaptive e_1 shape measure of the source as measured on the difference image[39].
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`
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

4.3.2 `DIAObject` Table

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

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[40]
parallax	float	mas	Parallax
pmParallaxCov	float[6]	various	Proper motion - parallax covariances.
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`.
lcPeriodic	float[6×32]		Periodic features extracted from light-curves using generalized Lomb-Scargle periodogram (Table 4, Richards et al. 2011)[41]

Continued on next page

[40]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.

[41]The exact features in use when LSST begins operations are likely to be different compared to the baseline described here. This is to be expected given the rapid pace of research in time domain astronomy. However, the *number* of computed features is unlikely to grow beyond the present estimate.

Table 2: `DIAObject` Table

Name	Type	Unit	Description
lcNonPeriodic	float[6×20]		Non-periodic features extracted from light-curves (Table 5, Richards et al. 2011)
nearbyObj	uint64[3]		Closest `Objects` in Level 2 database.
nearbyObjDist	float[3]	arcsec	Distances to `nearbyObj`.
nearbyObjLnL	float[3]		Natural log likelihood that the observed `DIAObject` is the same object as the nearby `Object`.
flags	bit[64]	bit	Flags

4.3.3 `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[42]

Continued on next page

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

Table 3: `SSObject` Table

Name	Type	Unit	Description
moidLon	double[2]	degrees	MOID longitudes.
H	float[6]	mag	Mean absolute magnitude, per band.
G	float[6]	mag	Fitted slope parameter, per band[43]
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 in LSST bands.

4.3.4 Precovery Measurements

When a new `DIASource` is detected, it's useful to perform PSF photometry at the location of the new source on images taken prior to discovery. These are colloquially know as *precovery measurements*[44]. 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 photometry on visits taken over the previous 30 days[45].

[43]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 using strong priors on slopes poorly constrained by the data.

[44]When Solar System objects are concerned, precovery has a slightly different meaning: predicting the positions of newly identified `SSObjects` on previously acquired visits, and associating with them the `DIASources` consistent with these predictions.

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

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 yet timely precovery of targets of special interest.

4.3.5 Reprocessing the Level 1 Data Set

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 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 would be far from optimal. The instrument will be better understood with time. Newer versions of LSST pipelines will improve detection and measurements on older data. Also, precovery photometry should optimally be performed for *all* objects, and not just a select few. 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 we perform the annual Level 2 data release productions. This will include all images taken since the start of survey operations, to the time when the data release 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 the reprocessing may take as long as ~ 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 *signifi-*

cant side-effects on downstream users. For example, all `DIASource` and `DIAObject` IDs will change. 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 may 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` lightcurve characterization).

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 exist to support the brokers who will use it locally to perform more advanced brokering[46].

Older versions of the Level 1 database will be archived following the same rules as for the Level 2 databases. 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 made available as bulk downloads. See § 5.5 for more detail.

4.4 Level 1 Image Products

4.4.1 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).

[46] 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).

4.4.2 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).

4.4.3 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 transient or fast moving objects (eg., asteroids) from the templates.

The input images may be further grouped by airmass and/or seeing[47]. 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.

4.5 Alerts to `DIASources`

4.5.1 Information Contained in Each Alert

For each detected `DIASource`, LSST will emit an "Event Alert" within 60 seconds of the end of visit (defined as the end of image readout from the LSST Camera). These alerts will be issued in `VOEvent` format[48], and should be readable by `VOEvent`-compliant clients.

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

- *alertID*: An ID uniquely identifying this alert. It can also be used to execute a query against the Level 1 database as it existed when this alert was issued
- *Level 1 database ID*: For example, `DR5L1`

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

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

- 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)

The items above are meant to represent the *information* transmitted with each alert; the content of the alert packet itself will be formatted to confirm to `VOEvent` (or other relevant) standard. Where the existing standard is inadequate for LSST needs, LSST will propose extensions and work with the community to reach a common solution.

With each alert, we attempt to include as much information known to LSST about the `DIASource` as possible, to minimize the need for follow-up database queries. This speeds up classification and decision making at the user end, and relaxes the requirements on the database on the Project end.

4.5.2 Receiving and Filtering the Alerts

Alerts will be transmitted in `VOEvent` format, using standard IVOA protocols (eg., VOEvent Transport Protocol; VTP[49]. As a very high rate of alerts is expected, approaching $\sim$ 2 million per night, we plan for public VOEvent Event Brokers[50] to be the primary end-points of LSST's event streams. End-users will use these brokers to classify and filter events for subsets fitting their science goals. End-users will *not* be able to subscribe to full, unfiltered, alert streams coming directly from LSST[51].

[49] VOEvent Transport Protocol is currently an IVOA Note, but we understand work is under way to finalize and bring it up to full IVOA Recommendation status.

[50] 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).

[51] This is due to finite network bandwidth available: for example, a 100 end-users subscribing to a $\sim$ 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.

To directly serve the end-users, LSST will provide a basic, limited capacity, alert filtering service. This service will run at the LSST U.S. Archive Center (at NCSA). It will let astronomers create simple filters that limit what alerts are ultimately forwarded to them[52]. 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., no cross-matches to 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 limited – in case of overwhelming interest, a TAC-like proposal process may be instituted.

[52]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.

4.6 Open Issues

What follows is a (non-exhaustive) list of issues, technical and scientific, that are still being discussed and where changes are possible. The estimate of the time by which a decision should be made is noted in parentheses.

- *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? (FDR)
- *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.). (FDR)
- *Should we associate alerts with external catalogs?.* LSST will have a copy of many external catalogs (or parts thereof), and could in principle provide limited positional association. Right now, it's up to the downstream brokers to perform such associations. (CONSTRUCTION)
- *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). (CONSTRUCTION)
- *What light-curve metric should we compute and provide with alerts?* We strive to compute general purpose metrics which will facilitate classification. We have currently baselined the Richards et al. (2011) feature set. (COMMISSIONING)
- *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. (COMMISSIONING)
- *Should the postage stamps provided with the alerts be binned, and by what factor?* The thought is that a human user may have an easier time performing a final by-eye check with larger postage stamps. (COMMISSIONING)

- *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. (OPERATIONS)

5 Level 2 Data Products

5.1 Overview

Level 2 data products result from direct image[53] 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` in Level 2 will be generated by associating and deblending the list of single-epoch source detections and the lists of sources detected on coadds. We plan to build coadds designed to maximize depth (*"deep coadds"*) and coadds designed to achieve a good combination of depth and seeing (*"best seeing coadds"*). We will also build a series of *short-period* (eg. yearly, or multi-year) coadds. The flux limit in deep coadds will be significantly fainter than in individual visits, and the best seeing coadds will help with deblending the detected sources. The short-period coadds are necessary to avoid missing faint objects showing long-term variability. These coadds will be built for all bands, as well as some combining multiple bands (*"multi-color coadds"*). ***Not all of these will be preserved*** after sources are detected and measured (see § 5.4.3 for details). We will provide a facility to regenerate their subsections as Level 3 tasks (§ 6).

The deblender will be run simultaneously on the catalog of peaks[54] detected in the coadds, the `DIAObject` catalog from the Level 1 database, and one or more external catalogs. It will use the knowledge of peak positions,

[53] As opposed to *difference image*, for Level 1.

[54] 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.

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 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[55].

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* (see § 5.2.3), and the flux measurements will be stored in the `ForcedSource` table.

5.2 Level 2 Data Processing

Figure 1 presents a high-level overview of the Level 2 data processing workflow[56]. Logically[57], 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.

[55] "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.

[56] Note that some LSST documents refer to *Data Release Processing*, which includes both Level 1 reprocessing (see § 4.3.5), and the Level 2 processing described here.

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

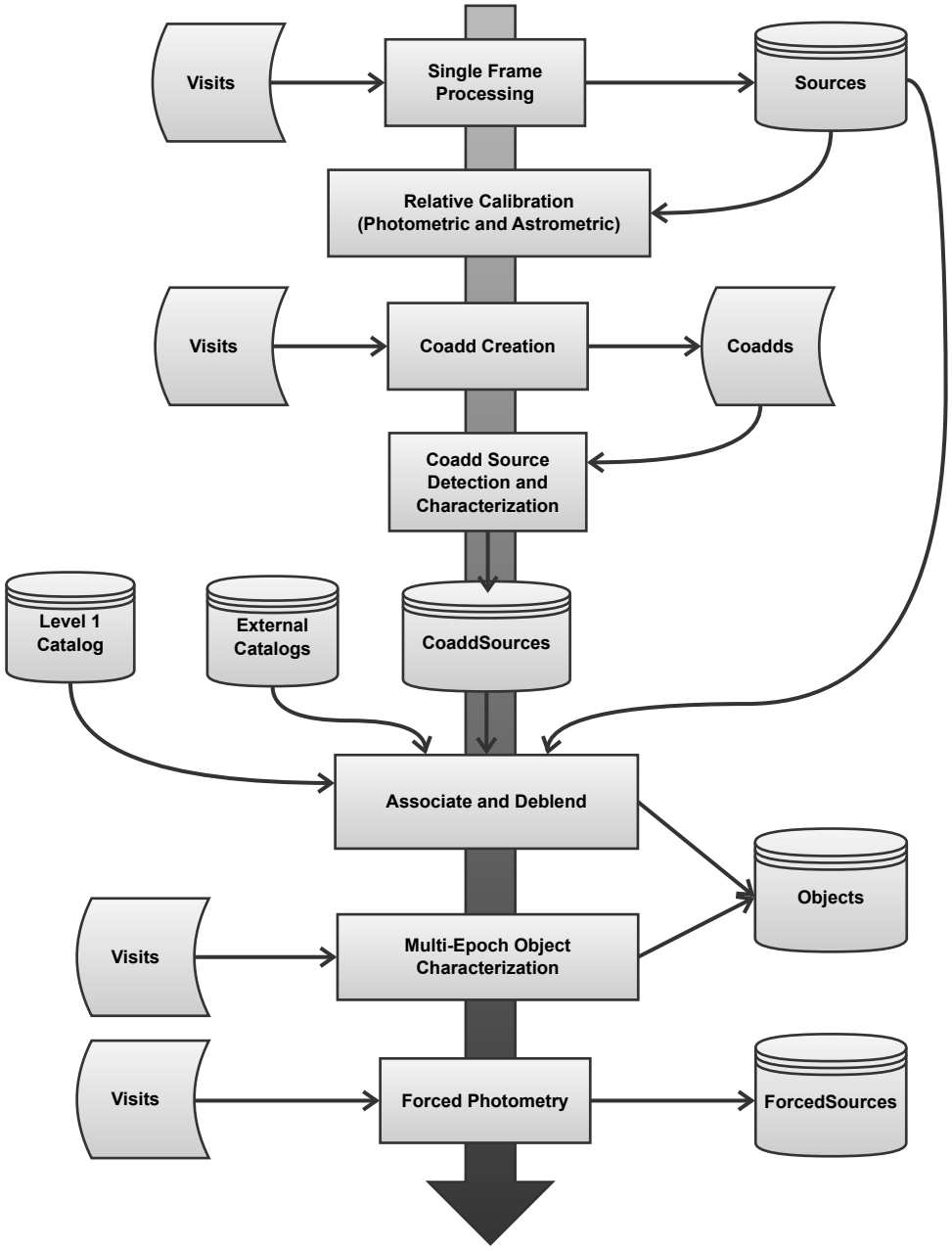

Figure 1: Level 2 Data Processing Overview

3. *Coadd creation*: Deep, seeing optimized, and short-period per-band coadds are created in *ugrizy* bands, as well as deeper, multi-color, coadds[58]. Transient sources (including Solar System objects, explosive transients, etc), will be rejected from the coadds. See § 5.4.3 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`[59].

5. *Association and deblending.* The next stage in the pipeline, which we will for simplicity just call *the deblender*, will synthesize a list of unique objects. In doing so it will consider the catalogs of `Sources` and `CoaddSources`, catalogs of `DIASources`, `DIAObjects` and `SSObjects` 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`[60].

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).

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

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

[60] 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 will be measured on every visit, with the position, motion, shape, and the deblending parameters characterized in the previous step kept fixed. This process of *forced photometry*, will result in the characterization of the light-curve for each object in the survey. The fluxes will be stored in the `ForcedSource` table.

5.2.1 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 § 5.3.1.

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 § 5.2.2.

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 (allowed to be 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[61].

- *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[62]. The

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

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

object is assumed not to move[63]. 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 covariance matrix, 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[64].

- *Centroids.* Centroids will be computed independently for each band using an algorithm similar to that employed by SDSS. Information from all[65] 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

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

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

[65] Whenever we say *all*, it should be understood that this does not preclude reasonable data quality cuts to exclude data that would otherwise degrade the measurement.

and *homogenized*, convolved to a canonical circular PSF[66]. The radii 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*[67] (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[68] of concentric, logarithmically spaced, PSF-homogenized, elliptical apertures, in standard seeing.

- *Variability characterization.* Two groups of parameters will be provided, designed to characterize periodic and aperiodic variability features (Richards et al. 2011). We caution that the exact features in use when LSST begins operations are likely to be different compared to the baseline described here; this is to be expected given the rapid pace of research in time domain astronomy. However, their *number* is unlikely to grow beyond the present estimate.

5.2.2 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 § 5.2, and may support other LSST science cases as seen fit by the users (for example, searches for objects whose shapes vary over time).

The following source properties are planned to be measured:

[66]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 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.

[67]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.

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

- *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.

- *Aperture surface brightness.* Aperture surface brightness will be computed in a variable number[69] 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[70], and will be better characterized by MultiFit (in the `Object` table).

5.2.3 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 single-visit 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.

5.2.4 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

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

[70] 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.

Clouds, and a number of globular clusters (among others).

LSST image processing and measurement software, although primarily designed to operate in non-crowded regions, is expected to perform well in areas of crowding. The current LSST applications development plan envisions making the deblender aware of Galactic longitude and latitude, and permitting it to use that information as a prior when deciding how to deblend objects. While not guaranteed to reach the accuracy or completeness of purpose-built crowded field photometry codes, we expect this approach will yield acceptable results even in areas of moderately high crowding.

Note that this discussion only pertains to processing of *direct images*. Crowding is not expected to significantly impact the quality of data products derived from *difference images* (i.e., Level 1).

5.3 The Level 2 Catalogs

This section presents the contents of key Level 2 catalog tables. As was the case for Level 1 (see § 4.3), here we present the *conceptual schemas* for the most important Level 2 tables (the `Object`, `Source`, and `ForcedSource` 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 named `band` in the `Source` table. In the as-built database, the views presented to the users will be appropriately denormalized for ease of use.

5.3.1 The `Object` Table

Table 4: Level 2 Catalog `Object` Table

Name	Type	Unit	Description
objectId	uint64		Unique object identifier

Continued on next page

Table 4: Level 2 Catalog `Object` Table

Name	Type	Unit	Description
parentObjectId	uint64		ID of the parent `Object` this object has been deblended from, if any.
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.
psParallax	float	mas	Point source model: Parallax.
psFlux	float[ugrizy]	nmgy	Point source model fluxes[71].
psCov	float[66]	various	Point-source model covariance matrix[72].
psLnL	float		Natural *log* likelihood of the observed data given the point source model.
bdRadec	double[2]	degrees	B+D model[73]: (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	B+D model: Integrated flux of the de Vaucouleurs component.

Continued on next page

[71] 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.

[72] Not all elements of the covariance matrix need to 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 ($\sim 1\%$).

[73] 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.

Table 4: Level 2 Catalog `Object` Table

Name	Type	Unit	Description
bdFluxD	float[ugrizy]	nmgy	B+D model: Integrated 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[74].
bdLnL	float		Natural *log* likelihood of the observed data given the bulge+disk model.
bdSamples	float[19][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.

Continued on next page

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

Table 4: Level 2 Catalog `Object` Table

Name	Type	Unit	Description
stdColorSigma	float[5]	mag	Uncertainty of `stdColor`.
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`.
E1	float[ugrizy]		Adaptive e_1 shape measure. See Bernstein & Jarvis (2002) for detailed discussion of all adaptive-moment related quantities[75].
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.

Continued on next page

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

Table 4: Level 2 Catalog `Object` Table

Name	Type	Unit	Description
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)
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/as^2	Mean surface brightness within an annulus[76].
apMeanSbSigma	float[6][`apN`]	nmgy/as^2	Standard deviation of `apMeanSb`.

Continued on next page

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

Table 4: Level 2 Catalog `Object` 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.
lcPeriodic	float[6×32]		Periodic features extracted from light-curves using generalized Lomb-Scargle periodogram (Table 4, Richards et al. 2011).
lcNonPeriodic	float[6×20]		Non-periodic features extracted from light-curves (Table 5, Richards et al. 2011).
photoZ	float[2×100]		Photometric redshift likelihood samples – pairs of (z, $logL$) – computed using a to-be-determined published and widely accepted algorithm at the time of LSST Commissioning.
flags	bit[128]	bit	Flags

5.3.2 `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).

Table 5: Level 2 Catalog `Source` Table

Name	Type	Unit	Description
sourceId	uint64		Unique source identifier[77]
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.
parentSourceId	uint64		ID of the parent `Source` this source has been deblended from, 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[78].
psLnL	float		Natural *log* likelihood of the observed data given the point source model.

Continued on next page

[77] 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.

[78] 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
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`.
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[64]	bit	Flags

5.3.3 `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.
flags	bit[8]	bit	Flags

5.4 Level 2 Image Products

5.4.1 Visit Images

Raw exposures, including individual snaps, and processed visit images will be made available for download as FITS files. They will be downloadable both through a human-friendly Science User Interface, as well as using 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[79].

5.4.2 Calibration Data

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

All auxiliary telescope data, both raw (images with spectra) and processed (calibrated spectra, derived atmosphere models), will be preserved and made available for download.

[79] Assuming identically performing software and hardware configuration.

5.4.3 Coadded Images

In course of Level 2 processing, multiple classes and numerous 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. 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 omitted). Transient sources (including Solar System objects, explosive transients, etc), will be removed. Care will be taken to preserve the astrophysical backgrounds[80].

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

- A set of *short-period* coadds. These will comprise of multiple (ugrizyM) sets of yearly and multi-year coadds. Each of these sets will be created using only a subset of the data, and otherwise share the characteristics of the deep coadds described above. These are designed to enable detection of long-term variable or moving[81] 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.

- A set of *best seeing* coadds. One deep coadd will be created for each of the *ugrizy* bands, using only the best seeing data (for example, using only the first quartile of the realized seeing distribution). These will be built to assist the deblending process. **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 and shapes of objects at "standard" seeing. ***We do not plan to keep and make these coadds available***. We will

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

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

retain and provide sufficient metadata for users to re-create them using Level 3 or other resources.

The exact details of which coadds to build, and which ones to keep, can change during Construction without affecting the processing system design as the most expensive operations (raw image input and warping) are constant in the number of coadds produced. The data management system design *is* sensitive to the total *number and size* of coadds to be *kept* – these are the relevant constraining variables.

To build the coadds, we currently plan to subdivide the sky into 12 overlapping[82] *tracts*, spanning approximately 75×72 degrees. The sky will be stereographically projected onto the tracts[83], and 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[84] 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**[85], 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.

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.

5.5 Data Release Availability and Retention Policies

Over 10 years of operations, LSST will produce eleven data releases: two for the first year of survey operations, and one every subsequent year. Each data

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

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

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

[85] 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.

release will include reprocessing of all data from the start of the survey, up to the cutoff date for that release.

The contents of data releases are expected to range from XX PB (DR1) to $\sim$ 70 PB for DR11 (this includes the raw images, retained coadds, and catalogs). Given that scale, it is not feasible to keep all data releases loaded and accessible at all times.

Instead, ***only the contents of the most recent data release, and the penultimate data release will be kept on fast storage and with catalogs loaded into the database***. Statistics collected by prior surveys (eg., SDSS) show that users nearly always prefer accessing the most recent data release, but sometimes may use the penultimate one (this is especially true just after the publication of a new data release). Older releases are used rarely.

To assist with data quality monitoring and assessment ***small, overlapping, samples of data from older releases will be kept loaded in the database***. The sample size is expected to be on order of $\sim 1 - 5\%$ of the data release data, with larger samples kept early on in the survey. The goal is to allow one to test how the reported characterization of the same data varies from release to release.

Older 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). Database software and data loading scripts will be provided for users who wish to set up a running copy of an older (or current) data release database on their systems.

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.

5.6 Open Issues

What follows is a (non-exhaustive) list of issues, technical and scientific, that are still being discussed and where changes are possible. The estimate of the time by which a decision should be made is noted in parentheses.

- *Is our approach to crowded field photometry satisfactory?* Is there another, given the budget and schedule constraints? (FDR)

- *What is the definition of a crowded field?* This is related to the question of when is LSST photometry allowed to deviate from SRD specs due to crowding. (FDR)

- *Are we going to provide additional characterization for large galaxies (isophote twists, bars, etc)?* How far should (and can) we go with "large" galaxy characterization. Measures like isophote twists, Gini coefficients, etc. This is now considered to be Level 3, but it's likely at least a portion of it belongs to Level 2 – assuming we're given a set of well-defined morphology measures. Input from the galaxies (and strong lensing?) Collaborations would be useful. (CONSTRUCTION)

- *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. (CONSTRUCTION)

- *What is the primary use-case for retained coadds?* These are the set of coadds (one per band) that we'll serve to the users. For example, should the primary purpose of deep coadds be the study of diffuse structures (in which case we may prefer to optimize depth at the cost of degrading the seeing), or is it to enable alternative studies of object shapes, where the resolution (and minimization of systematics) is paramount? This is very much related to the question above. (CONSTRUCTION)

- *Do we need both Kron and Petrosian fluxes?* These are intimately related to aperture fluxes, and given aperture fluxes in multiple apertures may be even derived from these. (OPERATIONS)

- *Do we need multiple estimates of object centroids?* The document currently assumes that the model-independent centroiding algorithm, the point source model fit, and the extended source model fit, don't necessarily need to agree on the location of the center of an object. This is true in general. However, it has been pointed out that when they do not agree, the associated flux measurement is meaningless anyway. (OPERATIONS)

6 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, and
3. Level 3 programming environment and framework

Many scientists' work may involve using two or all three of them in concert, but they can each be used independently. We describe each one of them in the subsections to follow.

6.1 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 and units as the corresponding L1/L2 columns.

There is no requirement that Level 3 data products (L3DPs) 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, measurement systems, etc. 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.

6.2 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. There may be other co-located but *not* project-owned,

resources available at the LSST DACs[86]; their use is beyond the scope of this document, except to note that reasonable provisions will be made to ensure it *is* possible to use them to process large quantities of LSST data.

Level 3 processing resources will, at least, include systems that can carry out traditional batch-style processing, probably similarly configured to those LSST will be using for the bulk of data release 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 § 6.3).

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.

6.3 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.

[86]For example, the U.S. DAC will be located at the National Petascale Facility building at NCSA, adjacent to the Blue Waters supercomputer.

These analyses could 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[87] 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. Early community engagement is therefore key to developing and maintaining these capabilities.

- 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.

[87] For example, BaBar.

6.4 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/frameworks 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.

7 Data Products for Special Programs

LSST Survey Specifications (LSST SRD, § 3.4) specify that 90% of LSST observing time will be spend executing the so-called "universal cadence". These observations will result in Level 1 and 2 data products described earlier in this document.

The remaining 10% of observing time will be devoted to special programs, obtaining improved coverage of interesting regions of observational parameter space. Examples include very deep ($r \sim 26$, per exposure) observations, observations with very short revisit times ($\sim$1 minute), and observations of "special" regions such as the Ecliptic, Galactic plane, and the Large and Small Magellanic Clouds. A third type of survey, micro-surveys, that would use about 1% of the time, may also be considered.

The details of these special programs or micro surveys are not yet defined[88]. Consequently, the specifics of their data products are left undefined at this time. Instead, we just specify the *constraints* on these data products, given the adopted Level 1/2/3 architecture. It is understood that no special program will be selected that does not fit these constraints[89]. This allows us to size and construct the data management system, without knowing the exact definition of these programs this far in advance.

Processing for special programs will make use of the same software stack and computing capabilities as the processing for universal cadence. The programs are expected to use no more than $\sim$10% of computational and storage capacity of the LSST data processing cluster. When special products include time domain event alerts, their processing shall generally be subject to the same latency requirements as Level 1 data products.

For simplicity of use and consistency, the data products for special programs will be stored in databases separate from the "main" (Level 1 and 2) databases. The system will, however, allow for simple federation with Level 1/2/3 data products (i.e., cross-queries and joins).

As a concrete example, a data product complement for a "deep drilling" field designed for supernova discovery and characterization may consist of: i) alerts to events discovered by differencing the science images against a special deep drilling template, ii) a Level 1-like database iii) one or more "nightly

[88]The initial complement is expected to be defined and selected no later than Science Verification.

[89]Or will come with additional, external, funding, capabilities, and/or expertise.

co-adds" (co-adds built using the data from the entire night), produced and made available within $\sim$ 24 hours, and iv) special deep templates, built using the best recently acquired seeing data, produced on a fortnightly cadence.

Note that the data rights and access rules apply just as they would for for Level 1/2/3. For example, while generated event alerts (if any) will be accessible world-wide, the image and catalog products will be restricted to users with LSST data rights.