LSST Data Products Definition (DRAFT)

Mario Jurić (`mjuric@lsst.org`)

with input from

T. Axelrod, A.C. Becker, J. Becla, J. Kantor, K-T Lim,
and R. Lupton

for LSST Data Management

April 15, 2013

Abstract

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

Contents

1 Introduction

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

The LSST will be operated in fully automated survey mode. The images acquired by the LSST Camera will be processed by LSST Data Management software to a) detect and characterize imaged astrophysical sources and b) detect and characterize changes in time in LSST-observed universe. The results of that processing will be reduced images, catalogs of detected objects and the measurements of their properties, and prompt alerts to "transients" – changes in astrophysical scenery discovered by differencing incoming images against older, deeper, images of the sky in the same direction (*templates*, see §2.3.4).

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

1.1 Level 1 and 2 Data Products

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

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

2. Analysis of science images, with the goal of detecting and characterizing astrophysical objects. Detection of faint galaxies on deep co-adds

[1]The LSST has adopted the terminology where single-epoch detections of astrophysical phenomena, of astrophysical *objects*, are called *sources*. Note that some other surveys call *detections* what we call *sources*, and use the name *sources* for what we call *objects*.

[2]`SSObject` used to be called call a "Moving Object". The name is potentially confusing, as high-proper motion stars are moving objects as well. A more accurate distinction is the one between objects in an out of the Solar System.

and their subsequent characterization is an example of this analysis. The results are **Level 2** data products. These products, released annually[3], will include catalogs of `Objects` (detections on deep co-adds) and `Sources`[4] (measurements on individual science images), as well as fully reprocessed Level 1 data products (see §2.3.8). In contrast to the Level 1 database, which is updated in real-time, the Level 2 databases are static and will not change after release.

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

2 Level 1 Data Products

2.1 Overview

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

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

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

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

[5]We here use the term *transient* loosely, having it include all sources detected in a difference image (eg., variable stars, asteroids) and not just explosive phenomena (eg., supernovae).

[6]The LSST SRD considers Solar System object orbits as a Level 2 data product (LSST-SRD, Sec 3.5). Nevertheless, to successfully differentiate between them and other types of transients they're functionally considered a part of Level 1.

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

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).

`DIASources` detected on visits taken at different times are associated with `DIAObjects`. `DIAObjects` represent the underlying astrophysical phenomenon detected and measured by individual `DIASources`. The association can be done in two different ways: by assuming the underlying phenomenon is an object within the Solar System moving on an orbit around the Sun[8], or by assuming the underlying phenomenon is distant enough to only exhibit small parallactic and proper motion[9]. The latter type of association is performed during difference image analysis right after the image has been acquired. The former is done at daytime by the Moving Objects Processing Software (`MOPS`), unless the `DIASource` is an apparition of an already known Solar System object ("`SSObjects`") in which case it's flagged as such during difference image analysis. Note that `DIASources` that are not immediately recognized as Solar System objects will be broadcast as transient alerts at the end of difference image analysis[10].

2.2 Level 1 Data Processing

2.2.1 Difference Image Analysis

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

1. A visit is acquired and the images reduced and combined to a single science image (cosmic ray rejection, instrumental signature removal[11],

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

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

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

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

combining of snaps[12], etc.).

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

3. The flux and shape[13] of the DIASource are measured on the difference image. The science image is force-photometered at the position of the `DIASource` to obtain a measure of the absolute flux. No deblending will be attempted.

4. The Level 1 database (see §2.3) is searched for a `DIAObject` or `SSObject` with which to positionally associate the observed `DIASource`[14]. If no match is found, a new `DIAObject` is created and the observed `DIASource` is associated to it.

5. If the `DIASource` has been associated to an `SSObject` (a known moving object), alert processing terminates here (see section 2.2.2 for how it continues)[15].

6. The `DIAObject` measurements are updated with new data. All affected columns are recomputed, including proper motions, centroids, light curves, etc.

7. The Level 2 database[16] is searched for one or more `Objects` positionally close to the `DIAObject`, out to some maximum radius[17]. The IDs of these `Objects` are recorded in the `DIAObject` record and provided in the transient alert.

8. A `VOEvent` is issued that includes: the name of the Level 1 database, the timestamp of when this database has been queried to issue this

[12] A visit consists of two, nominally 15 second, exposures, called *snaps*.

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

[14] The association algorithm will guarantee that a `DIASource` is associated with one and only one `DIAObject` or `SSObject`. The algorithm will take into account the parallax and proper or Keplerian motions, as well as the errors in estimated positions of `DIAObject`, `SSObject`, and `DIASource` to find the maximally likely match.

[15] TODO: We will probably emit an alert for asteroids as well; this needs to be added to the text

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

[17] Eg., a few arcseconds.

`VOEvent`, the `DIASource` ID, the `DIAObject` ID[18], 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 2.4 for a more complete enumeration.

9. Precovery forced photometry is performed on any difference image overlapping the position of the `DIAObject` taken within the past 30 days, and added to the database within 24 hours. No additional alerts are issued with the precovery photometry.

2.2.2 Solar System Object Processing

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

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

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

3. The collection of tracklets collected over the past 30 days is analyzed for those *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 `DIAObject` record. `DIAObjects` "orphaned" by this unlinking are deleted.[20].

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

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

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

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

2.3 The Level 1 database

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

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

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 (e.g., colors, the light curve) separate from those measurements for the galaxy.

- *An asteroid occulting a star.* If associated with the star on first apparition, the association would need to be dissolved when the source is recognized as an asteroid (perhaps even as early as a day later).

- *A supernova on top of a pair of blended galaxies.* It is not clear in general to which galaxy this `DIASource` would belong. That in itself is a research question.

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

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

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

Philosophically, the adopted 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.*

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

2.3.1 `DIASource` Table

This is a table[24] of sources detected at $SNR \geq 5$ on difference images (`DIASources`). On average, we expect ~ 2000 `DIASources` per visit (~ 2M per night).

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	uint128		Unique source identifier
ccdVisitId	uint64		Id. of CCD and visit where this source was measured

Continued on next page

[24]For this and other tables that follow a *conceptual schema* is presented that conveys *what* data will be recorded in the 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 presented here are normalizes (i.e., contain no redundant information). For example, since the band of observation can be found by joining a `DIASource` table to table with exposure metadata, there's no column for 'band' in `DIASource` table. In the as-built database, the views presented to the users will be appropriately denormalized for ease of use.

Table 1: `DIASource` Table

Name	Type	Unit	Description
diaObjectId	uint128		Id. of the `DIAObject` this source was associated with[25]
ssObjectId	uint64		Id. of the `SSObject` this source has been linked to[26]
midPointTai	double	time	Time of mid-exposure for this DIASource.
radec	double[2]	degrees	(α, δ)[27]
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.[28]
psFlux	float	nmgy[29]	Calibrated flux for point source model. Note this actually measures the flux *difference* between the template and the science image.
psFluxStdev	float	nmgy	Estimated uncertainty of `psFlux` (standard deviation of the likelihood)
psLnL	float		Natural *log* likelihood of the observed data given the point source model.

Continued on next page

[25] diaObjectId will be NULL if ssObjectId is not NULL

[26] ssObjectId will be NULL if diaObjectId is not NULL

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

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

[29] 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 §2.3.6 for details.

Table 1: `DIASource` Table

Name	Type	Unit	Description
trailFlux	float	nmgy	Calibrated flux for a trailed source model[30,31]. Note this actually measures the flux *difference* between the template and the science image.
trailLength	float	arcsec	Maximum likelihood fit of trail length[32,33].
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 science image centered at the centroid measured on the difference image (forced photometry flux)

Continued on next page

[30] A *Trailed Source Model* attempts to fit an 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.

[31] 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.

[32] 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.

[33] 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
fpFluxStdev	float	nmgy	Estimated uncertainty of `fpFlux`
fpSky	float	DN	Estimated sky background at the position (centroid) of the object.
fpSkyStdev	float	DN	Estimated uncertainty of `fpSky`
moments	float[5]	various	Adaptive first and second moments $(I_x, I_y, I_{xx}, I_{yy}, I_{xy})$, measured on the difference image.
momentsStdev	float[5]	various	Estimated uncertainty for each entry in `moments`.
extendedness	float		A measure of extendedness, computed using a combination of available moments and model fluxes or from a likelihood ratio of point/trailed source models (exact algorithm TBD). $extendedness = 1$ implies a high degree of confidence that the source is extended. $extendedness = 0$ implies a high degree of confidence that the source is point-like.
flags	bit[64]	bit	Flags

2.3.2 `DIAObject` Table

Table 2: `DIAObject` Table

Name	Type	Unit	Description
diaObjectId	uint128		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[34]
plx	float	mas	Parallax
pmPlxCov	float[6]	various	Proper motion - parallax covariances.
psFlux	float[ugrizy]	nmgy	Weighted mean point-source model magnitude.
psFluxErr	float[ugrizy]	nmgy	Uncertainty in estimate of `psFlux`
psFluxStdev	float[ugrizy]	nmgy	Width of the `psFlux` distribution.
fpFlux	float[ugruzy]	nmgy	Weighted mean forced photometry flux.
fpFluxErr	float[ugrizy]	nmgy	Uncertainty in estimate of `fpFlux`
fpFluxStdev	float[ugrizy]	nmgy	Width of the `fpFlux` distribution.
lsPeriod	float[ugrizy]	day	Period (the coordinate of the highest peak in Lomb-Scargle periodogram)
lsStdev	float[ugrizy]	day	Width of the peak at `lsPeriod`.
lsPower	float[ugrizy]		Power associated with `lsPeriod` peak.

Continued on next page

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

Table 2: `DIAObject` Table

Name	Type	Unit	Description
lcChar	float[$6 \times M$]		Light-curve characterization summary statistics (e.g., 2nd moments, etc.). The exact contents, and an appropriate value of N, are to be determined in consultation with time-domain experts.
nearbyObj	uint128[3]		Closest `Objects` in Level 2 database.
nearbyObjDist	float[3]	arcsec	Distances to `nearbyObj`.
flags	bit[64]	bit	Flags

2.3.3 `SSObject` Table

Table 3: `SSObject` Table

Name	Type	Unit	Description
ssObjectId	uint64		Unique identifier
oe	double[7]	various	Osculating orbital elements at epoch (q, e, i, Ω, ω, M_0, epoch)
oeCov	double[21]	various	Covariance matrix for `oe`
arc	float	days	Arc of observation.
orbFitChi2	float		χ^2 for the orbital elements fit.
nOrbFit	int16		Number of observations used in the fit.
MOID	float[2]	AU	Minimum orbit intersection distances[35]

Continued on next page

[35] `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[36]
hStdev	float[6]	mag	Uncertainty in estimate of H
gStdev	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.

2.3.4 Image Differencing Templates

Templates for difference image analysis will be created by co-adding 6-months to a year long groups of visits. The templates may be further binned by airmass and/or seeing[37]. 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.

2.3.5 Likelihoods vs. Posteriors

Unless noted otherwise, maximum likelihood values will be quoted for all fitted parameters (measurements). Together with covariances, these will allow

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

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

the end-user to apply whatever prior they deem appropriate when computing posteriors[38].

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[39]. 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 accordingly so as to minimize incorrect accidental usage. For example, a column named `gFlux` may be the expectation value of the g-band flux, while `gFluxML` may be the maximum likelihood value.

2.3.6 Fluxes and Magnitudes

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

We quote fluxes in units of “maggie”. A maggie, as introduced by SDSS, is a linear measure of flux. An object with flux of one maggie (integrated over the bandpass) has an AB magnitude of 0:

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

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

We realize that the large majority of users will want to work with magnitudes. For convenience, we plan to provide columns with (Pogson) magnitudes[40], where values with negative flux will evaluate to `NULL`. Similarly, we will provide columns with flux expressed in Jy (and its error estimates).

---

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

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

[40] These will most likely be implemented as “virtual” or “computed” columns

2.3.7 Precovery

When a new `DIASource` is detected, it's useful to perform forced photometry at the location of the new source on images taken prior to discovery, colloquially know as *"precovery"*[41]. Doing precovery in real time over all previously taken visits is too I/O intensive to be feasible. We therefore plan the following:

1. For all newly discovered objects, perform precovery forced photometry on visits taken over the previous 30 days[42].

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 (e.g., SNe), while the latter will provide an opportunity for more extensive immediate precovery of targets of special interest.

2.3.8 Annual Reprocessings

In what we've described so far, the Level 1 database is continually being added to as new images are taken and `DIASources` identified. Every time a new `DIASource` is associated to an existing `DIAObject`, the `DIAObject` record is updated to incorporate new information brought in by the `DIASource`. Once discovered and measured, the `DIASources` are never re-measured at the pixel level.

This is not optimal. Newer versions of LSST pipelines are likely to improve measurements on older data. Also, forced photometry should be performed on the position of the `DIAObject` on all pre-discovery images.

We therefore plan to reprocess all image differencing-derived data (the Level 1 database), at the same time as we perform the annual Level 2 data release productions. This will include all images taken since the start of observation, to the time when the DR production begins. The reprocessed

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

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

images will be processed with a single version of the image differencing and measurement software, resulting in a consistent data set.

As reprocessing is expected to take on order of ~ 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 steps. Once it does, the existing Level 1 database will be replaced with the new one, and all future alerts will refer to the reprocessed Level 1 database. Alerts for new sources "discovered" during data release processing and/or the catch-up process will *not* be issued.

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

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

To ease the transition, third parties (VO 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 (e.g., when writing event-broker queries, filter on position, not on `DIAObject` ID, if possible, etc.). A Level 1 database distribution service, allowing for bulk downloads of the reprocessed Level 1 database, will need to be established to support the brokers who will use it locally to perform more advanced brokering[43].

Older versions of the Level 1 database will be archived following the same rules as for the Level 2 databases. DR1, the most recent DR, and the

[43] 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).

one preceding the most recent one will be kept on disk and loaded into the database. Others will be archived to tape and available as bulk downloads.

2.3.9 Repeatability of Queries

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

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

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

or, more generally:

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

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

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

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

2.3.10 Uniqueness of IDs across database versions

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

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

2.4 Transient Alerts

2.4.1 Information Contained in Each Transient Alert

For each detected `DIASource`, LSST will emit a "Transient Alert" within 60 seconds of the end of exposure. These alerts will be issued in `VOEvent` format, and should be readable by `VOEvent`-compliant clients.

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

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

2.4.2 Using the Transient Alerts

We plan to broadcast information about transient alerts in `VOEvent` format, using standard IVOA protocols (e.g., VOEvent Transport Protocol; VTP). As a very high rate of alerts is expected, approaching ~ 2 million per night, we plan for public VOEvent Event Brokers[44] to be the primary end-points of LSST's VTP streams. End-users will use these brokers to classify and filter events on the stream for those fitting their science goals. End-users will *not* be able to subscribe to full, unfiltered, alert streams coming directly from LSST[45].

[44] 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 successors.

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

For the end-users, LSST will provide a basic, limited capacity, transient alert filtering capability. This service will run at the LSST archive center (at NCSA). It will let astronomers create simple filters that limit what `VOEvents` are ultimately forwarded to them[46]. 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 transients within two
# effective radii of a galaxy. This is for illustration
# only; the exact methods/members/APIs may change.

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

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

[46]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

2.5 Open Issues to be Closed in time for the Data Products Review (Late May 2013)

What follows is a (non-exhaustive) list of issues that are still being discussed and where changes are likely. Input on any of these will be appreciated.

- *What light-curve metric should we compute and provide with transient alerts?* We strive to compute general purpose metrics which will facilitate classification. We have not baselined any yet.
- *Don't use the term 'transient' to refer to all* ***`DIASources`***. This causes confusion and is inconsistent with the usual meaning of *transient* as impermanent astrophysical source, lasting only for a short time. The term *transient alert* will also need to be changed.
- *Can we, should we, and how will we measure proper motions on difference images?* This is a non-trivial task (need to distinguish between dipoles that are artifacts, and those due to proper motions), without a clear science driver (since high proper motion stars will be discoverable using Level 2 catalogs).
- *Is a fully up-to-date Level 1 database after each alert technically feasible?* If not, we will delay making the updated Level 1 database available until the end of night. Transient alerts will still be issued within 60 seconds.
- *Is a precovery service technically feasible?* Currently, it doesn't appear to be.
- *Is Level 1 database required to be relational?*. A no-SQL solution may be more appropriate given the use cases envisioned so far.
- *Will LSST provide a limited-capability event broker, as described in §2.4.2?* The SRD seems to demand it, but the opposing view is that it's not clear how useful it would be (and it's not useful, we shouldn't waste resources to provide it).
- *Should we broadcast alerts to Solar System objects?*. I think the answer is yes – otherwise we will need to devise a separate mechanism to distribute orbit updates. This needs to be added to the document.

- *Do we have to, and can we, use 128 bit integers for IDs?*. If 64 bit integers are provably sufficient, they will take up less space and be better supported (technologically).