Science Data Quality Assessment for the Large Synoptic Survey Telescope

Richard A. Shaw[a], Deborah Levine[b], Timothy Axelrod[c], Russ R. Laher[b], Vince G. Mannings[b], and the LSST Data Management Team
[a]National Optical Astronomy Observatory, 950 N. Cherry Avenue, Tucson, AZ 85719, USA; [b]IPAC/California Institute of Technology, M/S 100-22, Pasadena, CA 91125, USA; [c]LSST Corporation, 933 N. Cherry Avenue, Tucson, AZ 85721, USA

ABSTRACT

LSST will have a Science Data Quality Assessment (SDQA) subsystem for the assessment of the data products that will be produced during the course of a 10 yr survey. The LSST will produce unprecedented volumes of astronomical data as it surveys the accessible sky every few nights. The SDQA subsystem will enable comparisons of the science data with expectations from prior experience and models, and with established requirements for the survey. While analogous systems have been built for previous large astronomical surveys, SDQA for LSST must meet a unique combination of challenges. Chief among them will be the extraordinary data rate and volume, which restricts the bulk of the quality computations to the automated processing stages, as revisiting the pixels for a post-facto evaluation is prohibitively expensive. The identification of appropriate scientific metrics is driven by the breadth of the expected science, the scope of the time-domain survey, the need to tap the widest possible pool of scientific expertise, and the historical tendency of new quality metrics to be crafted and refined as experience grows. Prior experience suggests that contemplative, off-line quality analyses are essential to distilling new automated quality metrics, so the SDQA architecture must support integrability with a variety of custom and community-based tools, and be flexible to embrace evolving QA demands. Finally, the time-domain nature of LSST means every exposure may be useful for some scientific purpose, so the model of quality thresholds must be sufficiently rich to reflect the quality demands of diverse science aims.

Keywords: Data quality assessment, automated data analysis, quality metrics, sky surveys, software design

1. INTRODUCTION

The Large Synoptic Survey Telescope (LSST) will be an 8.4 m telescope capable of imaging the entire accessible sky every few nights[8]. The imaging camera includes a focal plane which will be populated with 189 CCDs dedicated to imaging a 9.6 deg^2 field of view in the *u*, *g*, *r*, *i*, *z*, and *y* passbands[8]. Each 4K × 4K CCD has 16 parallel amplifiers, which are capable of reading out the entire focal plane within 2 s, at an aggregate data rate of 3.2 GB/s, into temporary storage[5]. The planned operations model calls for acquiring paired 15 s exposures at every pointing, followed by a slew to the next sky position where the next pair of exposures will be taken, and so on as long as sky conditions permit[9]. The survey is expected to cover approximately 20,000 deg^2 of the southern sky at a sampling of 0.2" pixel^{-1}, with as many as a few thousand visits per position over the course of the 10-year survey. The LSST data management (DM) system is responsible for processing the 15 TB of raw data that will be obtained on average each night. The processing consists of multiple passes through the data at different epochs following the observation, in order to meet the science goals of the LSST Survey[7]. The DM system consists of highly parallel pipelines that will produce a wide variety of data products, including transient alerts, images, and catalogs, which must be ready for science analysis by the world-wide science community[6,9]. The raw and processed data products, including images and source catalogs, will accumulate in multiple archive centers at the unprecedented rate (for astronomy) of ~6 PB yr^{-1}, of which the majority will consist of compressed images[6]. The source catalogs are expected to contain 2×10^{10} objects by the end of the survey, with roughly equal numbers of stars and galaxies[9].

The demands on the DM system are considerable, and are driven primarily by the huge data rates and volumes, and by the high observing cadence. But the success of the DM system will be judged primarily on the quality of the data products, and only secondarily by its efficiency in managing the data flow. Quality management is multi-faceted[10], and

Software and Cyberinfrastructure for Astronomy, edited by Nicole M. Radziwill, Alan Bridger,
Proc. of SPIE Vol. 7740, 77400H · © 2010 SPIE · CCC code: 0277-786X/10/$18 · doi: 10.1117/12.857293

data product quality involves recognizing, evaluating, and tracking elements that can be controlled, and reporting on those that cannot. The full range of quality assurance within the LSST systems will be very broad indeed, and includes aspects of hardware system design and development, process definition and control, and operations (to name but a few). Here we focus on Science Data Quality Assessment (SDQA), which as a system collects, evaluates, and records information about the quality of raw and derived data products from a primarily scientific perspective. The customers for this system include observatory scientists and engineers, users of the science data, and the DM system itself. Among the biggest challenges for creating an SDQA subsystem are the vast range of scientific objectives, the ambitious quality goals, the complexity of the processing, the near certainty of changing requirements after the system is initially deployed, and the need for an extraordinary degree of automation given the high data volumes and observing cadence. We discuss LSST science data quality in the context of other major surveys in §2, and in §3 we describe an adaptive approach to SDQA and describe specific examples of how science data quality will be manifested in the DM system. We conclude in §4 with a look at system design elements that have emerged or will be explored in the near future.

2. SCIENCE DATA QUALITY ASSESSMENT IN IMAGING SURVEYS

Imaging surveys of large regions of the sky have historically provided the raw material for substantial advances in astronomy. Wide-area, single-epoch surveys such as the Palomar Sky Survey, 2MASS[12] and the Sloan Digital Sky Survey[14] continue to fuel advances on many fronts, while multi-epoch, generally smaller-area surveys such as MACHO[1], OGLE-III[13], SuperMACHO[2], and the Palomar Transient Factory[11] (PTF) explore the time domain. These extraordinary, public datasets enjoy very wide use for a variety of investigations, and there is every expectation that this high usage will continue for the foreseeable future. One of the key reasons for this success is the close attention that the project teams paid to science data quality, and the care with which that quality was demonstrated. The accuracies that were achieved in a number of areas are remarkable[4], and collectively these surveys set a very high standard for LSST.

The LSST survey shares much in common with these prior surveys, but there are some important differences to keep in mind as well. First, the time-domain nature of the LSST survey, like most other multi-epoch surveys, means that essentially all on-sky images will likely be useful for some key scientific purpose. Therefore the model for data quality must be sufficiently rich to express quality with respect to a wide diversity of scientific purposes. For example, the quality of images obtained in poor seeing conditions with partial cloud cover may be insufficient to contribute meaningfully to a deep-detection co-add, but they may be fully adequate for tracking an evolving supernova light curve or providing an optical identification of a gamma-ray burst event. Quality control mechanisms in single-epoch imaging surveys are often oriented toward identifying and excluding data that does not meet one or more quality criteria, whereas LSST will obtain data even in rather poor conditions with the full expectation that the quality of source measurements will be tagged appropriately. A second major difference with prior surveys is that individual LSST exposures are comparatively short (nominally 15 s), and cover a large area of sky. This means that images obtained during partial cirrus cloud cover will suffer from spatially variable grey extinction from the resolved clouds. (In other types of imaging programs, longer exposures of a smaller patch of sky tend to average out the variable extinction.) The LSST photometric calibration plan is to characterize and correct for the grey extinction, which will introduce complexity in the data quality assessment at the catalog level. Finally, alerts for transient and variable objects will be generated in near real-time (within 60 s of the end of a visit) based on difference image processing, which places extraordinary demands on rapid, automated image quality analysis at a scale never before attempted.

There are a number of objectives for the SDQA subsystem, but they share the unifying purpose of quantifying and recording the scientific quality of all data products that will be created by the LSST DM system. The development of this subsystem will enable a number of capabilities, some of which are secondary for SDQA per se, but are very important in the context of the LSST project. These include:

- During commissioning SDQA will play a key role in the assessment of whether the telescope and camera, and their respective subsystems, have met their design specifications. Quantities such as the size and shape of the point-spread function (PSF) and its variation across the focal plane, as well as the strength of ghost images and scattered light will be of obvious interest during this period.

- SDQA will provide a quantitative basis for evaluating the application of calibration reference data (flat-fields, linearity correction, reference catalogs, focal plane illumination model, photometric scale and zero-point, etc.).

- SDQA will provide diagnostic tools that, among other things, will facilitate diagnosing problems with hardware (e.g., CCD health and performance, and delivered image quality) and DM software. Analysis of pipeline stage quality problems with SDQA will inform decisions taken by down-stream pipelines and processes of whether to abort or otherwise alter their processing. If a world coordinate solution has unacceptably large errors, for example, there is little point in attempting to associate detected sources with entries in the object catalog.

- SDQA will be a primary means to measure progress with respect to global survey goals, such as the photometric depth achieved in stacked images in each band over the sky, the local and global astrometric accuracy, and temporal coverage.

The operating plan calls for multiple epochs of data processing[6] or *productions*, including alert production (within 60 s of a visit), moving object orbit determination (daily), calibration production (perhaps every fortnight), data release production (annually, with an additional data release in year one), and a final, grand production at conclusion of the survey one decade after the start of science operations. Each production has a particular scientific emphasis, and places different demands on SDQA. Beyond its primary goals, the SDQA system will facilitate answering the next question that normally occurs once a quality problem is identified, namely, what is the origin of the problem? Supporting a rapid "drill-down" capability for project staff members who are diagnosing problems is particularly important for LSST, given its high level of complexity and automation. If history on other projects is any guide, this capability will be especially handy during LSST's commissioning period.

3. AN ADAPTIVE APPROACH TO SCIENCE DATA QUALITY ASSESSMENT

The SDQA subsystem is currently under development, but its path to full realization will likely be different in flavor than that of other DM subsystems. This is in part because of the novelty and data-intensive nature of the LSST survey, but it also reflects historical patterns of prior challenging missions. That is, there is a great deal to learn about LSST (i.e., the performance of the hardware, including the as-delivered telescope and camera), the best strategies for data processing, and new science demands and opportunities that will undoubtedly emerge during operations. So while many aspects of the architecture can be specified up front, that architecture must be sufficiently flexible to allow for what might be called a natural evolution in requirements and technique that will continue through the end of the survey (and perhaps beyond). We intend to approach this problem by leveraging the collected technical experience gleaned from past surveys and other experiments, and the scientific expertise of the LSST user community. We will validate our system design over the course of multiple data challenges, and we will continuously refine SDQA throughout commissioning and operations to incorporate new ideas or refinements of implemented QA methods.

3.1 Leveraging Prior Experience

There are many aspects of science data quality that are common among astronomical imaging surveys; an incomplete list of categories and examples of specific tests for quality are given in Table 1. Most of these tests can be automated, although in practice most prior surveys depended upon human visual inspection to confirm the computed quality metrics. This approach will simply not scale to the high data rates and observing cadence of LSST so a substantial effort will need to be invested in highly reliable, automated techniques to assess science data quality.

Table 1. Common Elements of Science Data Quality Assessment for Imaging Surveys.

Category	Example Quality Measures
Image artifact flagging	Static bad pixels, cosmic rays, saturation, satellite trails, electronic cross-talk
Image background	Ghost images, scattered light, detector health, moonlight, fringing, sky glow
Delivered image quality	Size of the PSF, PSF shape (e.g., ellipticity), variation of the PSF with position in the focal plane
Astrometric fidelity and stability	Deviation of WCS solution from catalog sources; spatial variation of RMS deviations
Photometric fidelity and stability	Uniformity of photometric depth; dispersion about expected stellar locus in CMDs and color-color plots

There is a very large community of scientists who are part of the LSST project through Science Collaboration teams. The membership as of this writing was nearly 300 people, and many of them have had direct experience with the major surveys mentioned in §2. We are actively drawing upon this expertise, which though deep is also geographically distributed.

3.2 Testing and Simulation

The development of the LSST DM system is being organized around a series of Data Challenges[7] with the aim of evaluating the system design and algorithmic approaches at progressively higher levels of computational performance and scientific fidelity. Input data for the production runs are taken from two sources: the public images from the CFHT Legacy Survey, which are well matched in many scientific respects to the data expected from LSST, and from high fidelity image simulations which model the LSST telescope and camera designs as well as the anticipated environmental factors. These data challenges provide an excellent opportunity to perform white-box testing of most aspects of SDQA, apart from those that relate to the environmental conditions at the time of the observations, and to increase the likelihood of deploying a relatively robust SDQA subsystem when observatory commissioning begins. One of us (RRL) is a developer for the ongoing PTF survey, and is actively developing data quality tools that have high potential applicability for LSST. An obvious advantage is the ability to field-test highly automated science data quality tools, and to explore the scalability of user interface concepts.

3.3 Operations Context

If history is any guide, the start of observatory commissioning, followed by full-up science operations, will be an intense period where a great deal will be learned about the performance of the as-delivered telescope and camera. It will also be a period where various data and processing defects are identified (in part through SDQA functionality) and corrected. And it will be the first opportunity to calibrate the system and to establish quantitative benchmarks for stability and repeatability of the science data under a variety of operating conditions. For SDQA, it will mean setting thresholds for a variety of quality parameters that will define off-nominal conditions. It is also likely that the limits to data quality and scientific accuracy will be understood in much greater detail, and more subtle qualities of the instrumental signature will become apparent with careful analysis. Once new characterizations or calibrations are built it is often straightforward to identify metrics that quantify the science data quality with respect to the new analysis, which can then be folded into the automated SDQA system. Such a process was undoubtedly at work, for instance, in the quality assessment of the SDSS photometric calibration[3] where the concise characterization of basic color-color diagrams was introduced well along in the project lifecycle. Thus, it is important to recognize the value of supporting and facilitating this phase of contemplative analysis, which is inevitably carried out by scientists or engineers with common, third-party software tools, operating on a variety of science and engineering data. It is likewise important to embrace this gradual migration and distillation of knowledge and technique from the user's workbench to the SDQA subsystem, and to fold this process into the subsystem design.

4. TOWARD A SYSTEM DESIGN

As described above, there are two broad, complementary activities that will be enabled with the SDQA subsystem: the automated computation, persistence, and flagging (via *thresholds*) of off-nominal conditions and quality attributes (i.e.: *ratings*) of pre-defined quantities (i.e.: *metrics*); and support for contemplative, post-facto assessment of science quality attributes by scientists and technical staff using a variety of custom and third-party tools operating on project data. The first category applies to quality attributes that easily lend themselves to simple parameterizations, or that are essential for time-critical assessments, or for which there is a good deal of experience with the generated data. The second category applies to long-term trend analyses; uncovering subtle problems with instrument signature removal, data calibration, or the processing system; and for investigating processing or quality anomalies. Experience with prior imaging surveys suggests that the contemplative, post-facto analyses will over time be distilled into concise and revealing quality metrics (including thresholds), which will then be folded back into the automated SDQA processing, making the DM system more robust. In fact, the high data rates and volumes place such a high value on efficiency and automation that this evolution is essential to project success.

In the end the SDQA system will enable the comparison of measured properties of the science data with expectations from prior experience or expectations from models. It will also enable the comparison of measured properties of the data

(as described in the science database) with established requirements for the scientific and technical performance of the observatory and the survey. During the survey the SDQA subsystem must enable comparisons of quality parameters between two or more productions. This capability is essential for validating the quality of new calibration reference files, or for validating new or updated software. While software validation is not strictly speaking a responsibility of the SDQA subsystem, it is an enabler of these kinds of quality processes.

4.1 Components of SDQA

The characterization of SDQA as a *subsystem* perhaps deserves some explanation. It is prohibitively expensive to process large quantities of data at the pixel level solely for the purpose of quality evaluation. Thus, all measures of quality at the pixel level are implemented in the individual pipeline stages, and the ratings are passed along as metadata in the production. In this sense many tasks related to science data quality are delegated to the *pipeline stages*, even though the responsibility for science data quality per se is managed at a higher level. This is in part a reflection of the high degree of data-parallelism in the architecture of the DM system. This parallelism also creates a need to monitor data quality across parallel threads—i.e., understanding data quality on the spatial scale of the focal plane requires integrating information for all processing of data segments (and, as seems likely, from environmental telemetry) from a single visit. Tracking and persisting the SDQA ratings from the parallel threads is the responsibility of the DM *middleware*, while the SDQA subsystem software appropriately assumes spatial and temporal levels of aggregation. The ultimate destination for the SDQA ratings and related metadata is the *science database*, a portion of which is described by schema that relate directly to the requirements of the SDQA subsystem. Finally, *various custom or third-party tools* in the form of user applications (including graphical user interfaces) must be included within SDQA in order to support the human level quality analysis described above.

4.2 Metrics

As described above, metrics are measures of interest in characterizing data quality, and when computed on real data can be compared in a highly automated way to expectations using (often scalar) thresholds. A list of specific metrics for quantifying LSST data quality is being compiled: see the growing list of metrics definitions for the third Data Challenge (DC3) on the DM project wiki, which is largely based on the collective experience of the LSST DM development team. The actual evaluation of quality can in principal involve ratings of multiple metrics, which we will explore in the future, possibly using advanced artificial intelligence techniques such as neural networks. The specific metrics that are being used for the current data challenge (DC3b) are heavily weighted toward diagnostics that verify the proper functioning of the DM pipeline stages. Metrics such as those related to the background level and uniformity over the focal plane help to assess the bias subtraction, flat-fielding, and fringe removal. When operations begin, these metrics are more likely to be used in monitoring the health and performance of the detectors, or to detect anomalous scattered light. Some of the current metrics are related to image artifact detection, which will be critical in the context of vetting transient object detections, where broadcasting an alert to the world-wide community warrants a careful, if brief, quality assessment. As DC3b progresses, we will be able to evaluate quality with respect to metrics more traditionally tied to science quality of imaging surveys, such as photometric and astrometric accuracy and stability.

4.3 Tools for Detailed Quality Analysis

The more complex, data-intensive experiments, such as large sky surveys, NASA astrophysics missions, etc., generally require a good deal of analysis to fully characterize and remove the instrumental signatures, and to characterize the scientific quality of the data. Much of this analysis is highly exploratory in nature, and involves the analysis of science and engineering data and metadata from a variety of sources. Often various algorithmic approaches are explored and refined, until an acceptable measure of data quality is achieved. We anticipate a similar need during the LSST commissioning and operations phases, but given the complexity of the data system it is important to address the requirements for such analyses in the DM system architecture. Figure 1 illustrates a possible model for user interaction with LSST data (including SDQA metadata). This model of activity explicitly recognizes that

1. When possible, users prefer to use software tools that are familiar to them.
2. Users may draw upon information beyond the contents of the LSST science database, possibly by applying transformations to data in the LSST science catalogs, or data drawn from the Virtual Observatory.
3. Users will often construct scripts or other programs to conduct their analysis.
4. Data stores, standard tools, and user programs need to communicate results dynamically.

We expect most users who wish to make use of algorithms or pipeline stages from the LSST code base will do so using Python, but that third-party tools such as DS9, TOPCAT. IRAF tasks, etc. will be used as well. This puts a premium on seamless communication of these software components and distributed data stores, for which SAMP appears to be a viable enabling technology and which is currently an International Virtual Observatory Alliance (IVOA) recommendation.

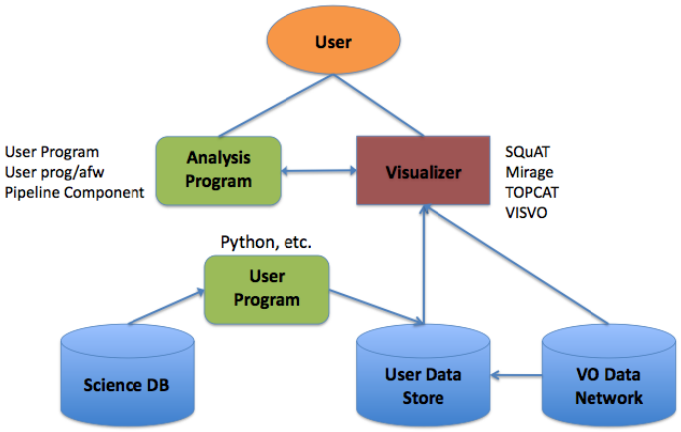

Figure 1. Schematic representation of potential interactions between users doing data quality analysis and various data stores, analysis software, and community utilities such as visualization tools. Connected pathways indicate data flow (arrows) or some form of bi-directional interaction (lines).

4.4 User Interface

A key requirement for SDQA is to provide for rapid identification of quality problems, along with enabling science and engineering staff to diagnose problems once they arise. While community based tools will no doubt play a part in supporting that effort, it is likely that one or more customized user interfaces (UIs) will be built for this purpose. Figure 2 illustrates the case of how a UI can facilitate monitoring of quality across the focal plane of the LSST camera. Bear in mind that the ability to assimilate information and present it in a comprehensible way to a human is not an inconsiderable challenge, given that LSST will generate 3.2 billion-pixel images at the rate of a few per minute. Thus, the ability to summarize, coupled with the ability to rapidly drill down to appropriate levels of detail will be critical in any UI-based tool. Figure 3 illustrates one browsing tool that was developed for quality assessment for the PTF survey. Practical experience with ongoing surveys is extremely helpful in exploring scalable user interface techniques for LSST.

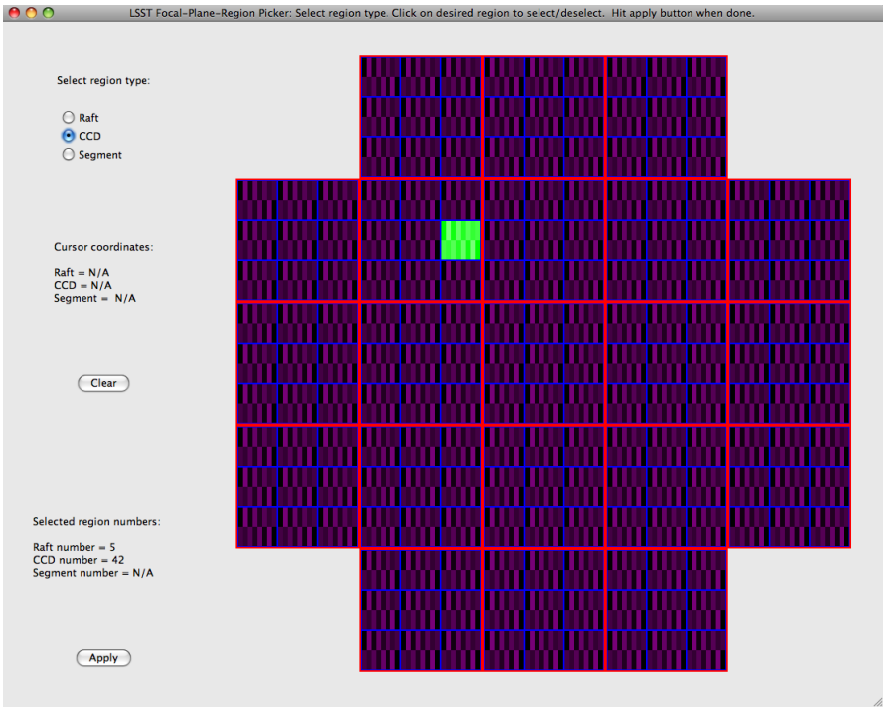

Figure 2. Mock-up of an SDQA user interface for visualizing the variation of quality ratings across the focal plane array. Areas of the FPA that exceed pre-defined thresholds for a selected quality metric are highlighted (green box) and summary information can be displayed. Drill-down to greater levels of detail could be enabled through mouse actions. This particular UI has potential utility in the operations environment, and combined with other environmental information (e.g., seeing monitors) could be useful for identifying detector problems, PSF anomalies, or light cirrus cloud cover.

Figure 3 Screen-shot of a data quality strip chart (quality ratings as a function of time) for the PTF project.

ACKNOWLEDGEMENTS

The LSST development work is the result of efforts by the LSST collaboration of scientist, engineers, technicians, managers as well as the study work contracted to several outside entities. This team of dedicated and recognized experts in their field is what makes the LSST project a success. At the 2008 annual LSST all hands meeting there were 160 people that participated from the project team and Science Collaborations.

LSST is a public-private partnership. Funding for design and development activity comes from the National Science Foundation, private donations, grants to universities, and in-kind support at Department of Energy laboratories and other LSSTC Institutional Members. This work is supported by in part the National Science Foundation under Scientific Program Order No. 9 (AST-0551161) and Scientific Program Order No. 1 (AST-0244680) through Cooperative Agreement AST-0132798. Portions of this work are supported by the Department of Energy under contract DE-AC02-76SF00515 with the Stanford Linear Accelerator Center, contract DE-AC02-98CH10886 with Brookhaven National Laboratory, and contract DE-AC52-07NA27344 with Lawrence Livermore National Laboratory. Additional funding comes from private donations, grants to universities, and in-kind support at Department of Energy laboratories and other LSSTC Institutional Members.

REFERENCES

[1] Alcock, C., et al. “The MACHO Project: Microlensing Detection Efficiency,” ApJS, 136, 439-462 (2001).
[2] Becker, Andrew C., et al. “The SuperMACHO Microlensing Survey,” Proc. IAU Symp. 225, 357-362 (2004).
[3] Ivezic, Z., et al. “SDSS Data Quality Management and Photometric Quality Assessment,” Astron. Nachr., 325, 583-589 (2004).
[4] Ivezic, Z., et al. “Sloan Digital Sky Survey Standard Star Catalog for Stripe 82: The Dawn of Industrial 1% Optical Photometry,” AJ, 134, 973-998 (2007).
[5] Kahn, S., et al. "Design and Development of the 3.2 Gigapixel Camera for the Large Synoptic Survey Telescope," Proc. SPIE 7733, in press (2010).
[6] Kantor, Jeffrey P., and Axelrod, Timothy “An Overview of the LSST Data Management System,” SPIE 7740-60, in press (2010).
[7] Kantor, Jeffrey P., “The Large Synoptic Survey Telescope Data Challenges,” Proc. SPIE 7740-61, in press (2010).
[8] Krabbendam, Victor L., and Sweeney, Donald “The Large Synoptic Survey Telescope Preliminary Design,” Proc. SPIE 7733-9, in press (2010).
[9] LSST Science Collaborations and LSST Project, [LSST Science Book], (Version 2.0; Tucson: LSST Corp.), available from arXiv:0912.0201 (2009).
[10] Radziwill, Nicole M., “Quality Management in Astronomical Software and Data Systems,” ASP Conf. Ser. 376, 363-372 (2007).
[11] Rau, Arne, et al. “Exploring the Optical Transient Sky with the Palomar Transient Factory,” PASP, 121, 1334-1351 (2009).
[12] Skrutskie, M. F., et al. “The Two Micron All Sky Survey (2MASS),” AJ, 131, 1163-1183 (2006).
[13] Udalski, A., et al. “The Optical Gravitational Lensing Experiment. OGLE-III Photometric Maps of the Large Magellanic Cloud,” AcA, 58, 89-102 (2008).
[14] York, Donald G., et al. “The Sloan Digital Sky Survey: Technical Summary,” AJ, 120, 1579-1787 (2000).