Draft version July 9, 2021
Typeset using LaTeX **modern** style in AASTeX62

Overview of LSST Data Management

William O'Mullane,[1] Leanne P. Guy,[1] Yusra AlSayyad,[2] Colin T. Slater,[3] and John D. Swinbank[3,2]

[1]*LSST Project Office, 950 N. Cherry Ave., Tucson, AZ 85719, USA*
[2]*Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544, USA*
[3]*University of Washington, Dept. of Astronomy, Box 351580, Seattle, WA 98195, USA*

(Dated: July 9, 2021)

ABSTRACT

Vera C. Rubin Observatory Data Management (DM) subsystem is one of four subsystems.

overview ...

1. INTRODUCTION

Within Rubin Observatory DM was tasked to stand up operable, maintainable, quality services to deliver high-quality LSST data products for science and education, all on time and within reasonable cost. DM is responsible for provided the tools necessary to take the bits generated by the telescope and turn them in to science ready products.

See also The Ruben Observatory Data Management System: Jurić et al. (2017)

1.1. *Science Drivers*

The astronomical size and complexity of the expected Rubin data drives many of the architectural choices made for the DM system. Teh following table highlights some of the key numbers that have influenced choices in DM.

Table 1. Rubin Key Numbers driving DM architectural choices

Parameters	**Number**	**Unit**
N Objects	40 billion	–
N Alerts per image	10 000	–
N Alerts per night	10 million	–

2. ORGANISATION OF DATA MANAGEMENT

The Organisation and management of DM is covered in detail in O'Mullane et al. (2018). As shown in Figure 1 DM Management is aligned mainly along the Work

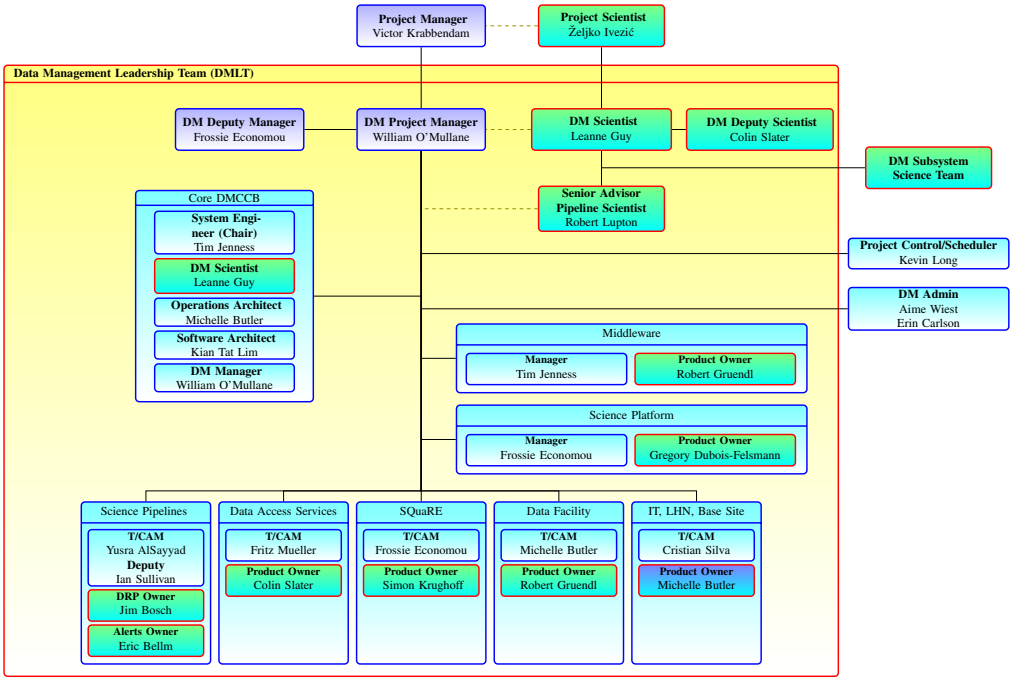

Figure 1. Org chart for Rubin DM from O'Mullane et al. (2018)

Breakdown Structure (WBS) of the project. It was found that a strict adherence to the WBS structure led to some products not being the responsibility's of a single manager when they spanned the organisation. To address this two cross cutting teams were identified to take care of the Science Platform and middleware. Each of these areas was then assigned to a manager to ensure its delivery - these managers were requested resources across the subsystem as needed.

Also of note in Figure 1 are the product owners. To ensure a single voice toward developers the product owner interprets requirements and sets priorities for the project. This also involves being the point of contact for any other stakeholders and incorporating their needs or wishes in the system. Interpretation of requirements is always difficult on large projects like Rubin observatory, they exist over a long period of time, were often written by people no longer on the project, and frequently are not easily verifiable. Hence another important roll of the product owner is in defining the verification tests for the requirements. Tests give a very concrete interpretation of the requirement. Verification covers all subsystems, software and data see PST (2020). DM will be verified and validated as part of System verification and validation Selvy et al. (2014) Bauer (2020)

2.1. *Open development process*

From the outset DM was seen as a large scientific software project. Agile methodologies Larman & Basili (2003) are particularly suited to the uncertainties of a science project and a cyclical approach to software development, with a period of six months

was adopted early on. A set of Epics corresponding to major pieces of work are defined at the beginning of each cycle. Tickets to track the work are created in Jira.

All code, and in fact documents, are kept under continuous integration using a mixture of Jenkins and GitHub Actions. Everything it under an open source licensed (mainly GPL) and available openly on GitHub.com. For the pipelines traditional releases are made each six months. Within SQuaRE services are released as needed and continuously deployed (currently with ArgoCD).

This is a large NSF funded project and so must still adhere to a more waterfall style of reporting. Milestones for major functionality, tied to major project milestones, were laid out and tracked in the usual manner. The DM approach to the Earned Value Management System (EVMS) used through Rubin construction is shown in Becla et al. (2016).

2.2. *Mode of work*

The DM team is distributed in several centers across the continental US as well as Chile and France. A strong set of guidelines developer.lsst.io was introduced early on to help homogenize modes of work e.g. dealing with tickets, naming github branches, merges, code style etc. It has functioned as a distributed organisation from the beginning, which probably help it weather the COVID-19 pandemic reasonably well. The Technical Account Managers (T/CAM), as shown in Figure 1, have a large degree of autonomy to deliver their software products. The Data Management Leadership Team (DMLT) comprises the managers and product owners and has a brief (30 minute) weekly meeting to set direction and raise issues (on Mondays). There is a longer multi day meeting three or four times a year some of which were physical get togethers, at least before the pandemic. The Managers have a standup meeting on Thursdays to work out any blockages or anticipated issues between the teams. Each team has its own regular meetings and discussions.

There is a mature decision making process where, in general, decisions are made at the lowest level possible within the team i.e. at the level of the individual developer where practical. This is enshrined the empowerment section of the guide. When this is not possible, decision making is escalated through the hierarchy using the Request for Comments (RFC) mechanism. DM captures decision making in technical notes (the DMTN series) or formal documents (the LDM series).

2.3. *Relationship to other subsystems*

As already mentioned all of Rubin interacts with System Engineering We take images from the LSST camera: Kahn et al. (2010)

We are commanded and listen to the Telescope and site software Gressler et al. (2014)

3. ARCHITECTURE, DATA TRANSMISSION AND ACCESS

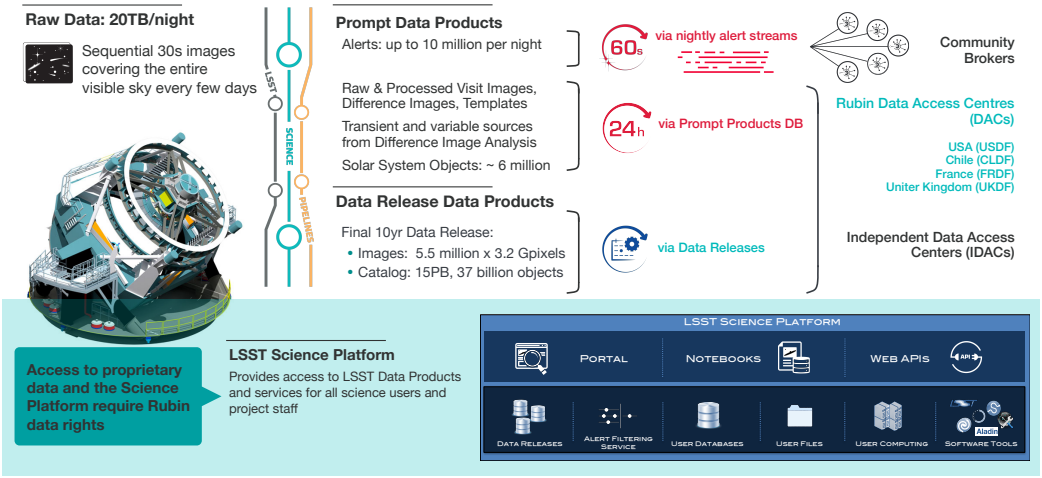

Figure 2. Overview of data managent from the telescope to the user.

DM spans multiple locations with processing occurring at the USDF (SLAC), FrDF (IN2P3) and UK (ROE).

Show network diagram and architecture diagrams for alerts/DRP. Mention Data Access doc Blum & the Rubin Operations Team (2020) and setting up of DACs in Chile and USA.

3.1. *Architecture*

3.2. *Networks*

3.3. *Data Access*

As shown in Figure 2 there are several kinds of Rubin data - mostly they are accessed via the science platform or a community broker. In addition some products may be accessed via and independent data access center. We will provide a few more details in this section.

3.3.1. *Data Access Centers*

3.3.2. *Alerts and Brokers*

Mention community alert brokers.

4. DM PRODUCTS

The DM products are not data as many may think, rather the products are software and services to produce those products. In operations DM will not exist though an analogous organization called Data Production will spring into existence under similar leadership.

The high level list of DM products is given in Figure 4, as may be seen in the figure we consider software, services and infrastructure as our categories of products.

It must be remarked that these products grew organically to some extent in a less than satisfactory manner. As mentioned earlier some teams worked within their WBS area and produced planning and products without necessarily paying a lot of attention

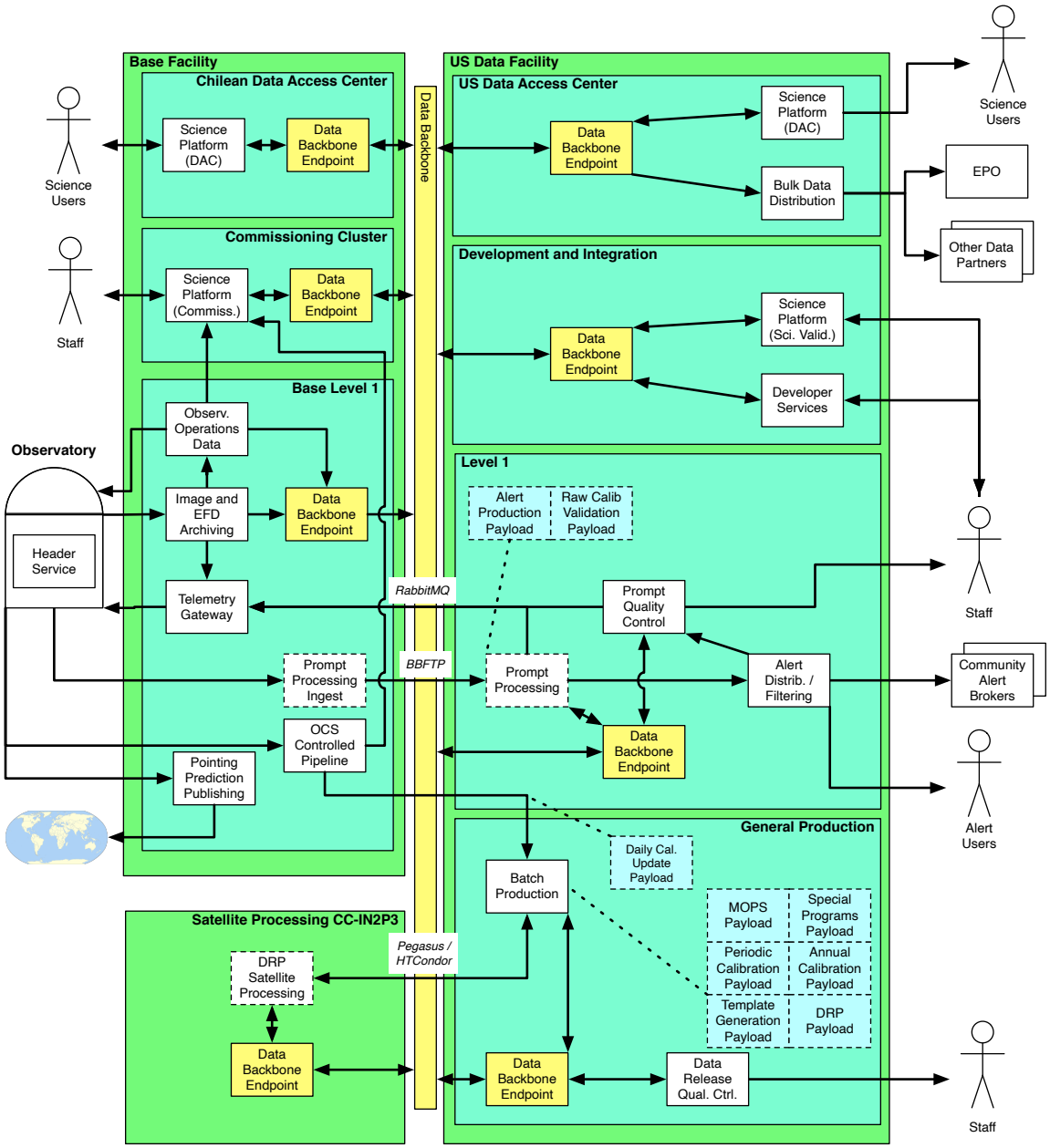

Figure 3. Rubin DM architecture diagram Lim et al. (2018)

to other WBS elements. Hence we have some services which are really deployments of software produced by another team e.g Prompt Services and Prompt Software. But we do not always have a service for a piece of software though it may be web accessible and look like a service e.g. QC Products. Some products are discussed in more below.

4.1. *Commissioning Software Products*

A detailed section describing what we delivered and used to process the commissioning data. This section will probably not be written until commissioning.

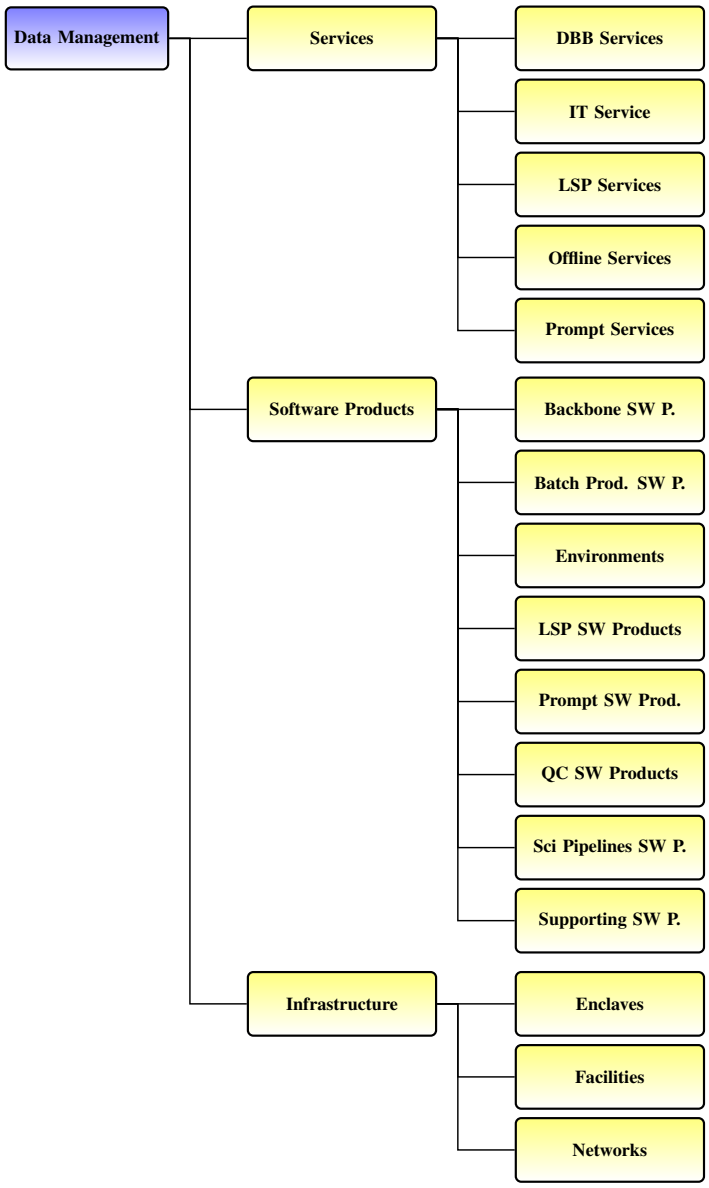

Figure 4. Rubin DM product tree

4.2. *Anticipated Data Release 1 Software Products*

A short section describing what we anticipate delivering for DR1 in addition to what was delivered in commissioning. This section will probably not be written until after commissioning.

4.3. *Science Platform*

The science platform came into existence as a concept around the end of 2016, there had always been requirements to allow near the data processing and to bring the code

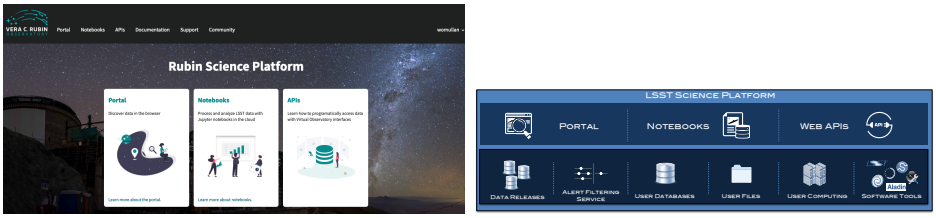

Figure 5. Rubin Science platform as released in 2021(left) and components(right)

to the data but no concrete idea of how to do this. Of course Ipython notebooks were an obvious tool for a python based project but the advent of Jupyter Hub (2015) lead to a more client server notion much more online with bringing code to data. The first version of the science platform visionJurić et al. (2017) was created in 2017.

4.4. *Science Pipelines*

The science pipelines which produce the prompt and data release products are of course the most identifiable software product from DM. The detailed approach to Science Pipelines is covered in Jenness (2020).

4.5. *Services*

4.6. *Infrastructure*

5. DATA PRODUCTS

The LSST data collected by Rubin Observatory is automatically processed by the LSST science pipelines, as described in (§ 4) to produce the LSST data products. The data products have been designed to enable the vast majority of LSST science, without the need to access the raw pixels or for users to reprocess the data. (it would be a gargantuan effort if not and totally unscalable). Some example science cases where pixel access or a reprocessing of the data might be warranted are, subtracting a different background (LSB science), reprocessing a small fraction of images to develop the systematics budget for weak lensing studies (DESC). (others?). This section provides a high-level overview of the LSST data products.

5.1. *Types of Data Product*

LSST produces three types of data products; images, catalogs and alerts.

`Images` — pricessed visit images (PVI) are images that have been corrected for instrumental effects and photometrically and astrometrically calibrated. raw single visit images, calibrated processed visit images (PVI), coadd images, cutouts (postage stamps)

`catalogs` — DR includes Object, Source, DIASource, DIAObject,

`alerts` — A composite data product that includes image cutouts (postage stamps) and extracts of catalog data. Alerts packets are distributed via the alert distribution

system (§ ref), one alert for each object that has changed in brightness or position on the sky.

5.2. *Categories of Data Product*

LSST defines three main categories of data products to be served by Rubin. The different categories are designed to enable different types of science. Each category of data product may comprise any or all of the data product types described in § 5.2.

5.2.1. *Prompt data products*

Prompt data products are designed to enable time domain science, the rapid discovery, characterization and follow up of objects that have been observed to change in position or brightness on the sky. *Add in a list of science cases that will be enabled on the various time scales* These data products are fully processed single visit images, difference images, and the catalogs produced by difference image analysis (DIA) (sec ref to software products). DIA outputs consist of, the sources detected in difference images (DIASources), the astrophysical objects that the sources are associated to (DIAObjects), characterizations of hitherto identified Solar System objects (SSObject), and discoveries of new Solar System objects.

Prompt data products are the result of nightly processing. Prompt data products are all based on difference imaging, and as such require transient-free templates to exist for each pointing and filter. The production of templates Prompt data products are release on a continual and ongoing basis. Two latencies, 60s for alerts and 24hrs for the catalogs. Data on likely optical transients, will be released publicly with a latency of at most 60s.

They are generated continuously every observing night, including both alerts to objects that have changed brightness or position, which are released with 60-second latency, and other catalog and image data products that are released with 24-hour latency. Prompt image data products include:

Image data products—PVIs,

Catalog data products—DIASource, DIAObject catalogs,

Alerts—

5.2.2. *Data Release data products*

A Data Release (DR) is specific, fixed `snapshots` of the data at a given time. Data Releases are made periodically and that can be used and unambiguously referenced in published analyses. The catalogs that form the data release will include an extensive list of quantities measured on sources detected in images and enable a variety of science analyses without the need for users to access or reprocess the images These data products will be made available as part of an LSST Data Release (§ **??**) as the result of coherent processings of the entire science data set to date. These will

include calibrated images, measurements of positions, fluxes, and shapes, variability information such as orbital parameters for moving objects, and an appropriate compact description of light curves. The Data Release data products will include a uniform reprocessing of the difference imaging-based Prompt data products.

5.2.3. *User Generated data products*

User Generated data products data products will originate entirely from the community, including project teams. These will be created and stored using suitable Application Programming Interfaces (APIs) that will be provided by the LSST Data Management System. The system will allow the science teams to use the full power of the Rubin database systems and Science Platform for the storage, access, and analysis of their results. It will provide for users and groups to maintain access control over the data products they create, enabling them to have limited distribution or to be shared with the entire LSST community.

The Rubin Science Platform (§ **??**) will allow for the creation of User Generated data products and will enable science cases that greatly benefit from co-location of user processing and/or data within the LSST Archive Center.

The first two, `Prompt` and `Data Release` data products are produced and delivered by the DM system described in this paper. The third, `User Generated` data products are produced by the Rubin Science Community using the `Prompt` and `Data Release` together possibly with data from other surveys.

The data product categories are outlined in Dubois-Felsmann et al. (2018)

In operations Data Production will use the the software outlined in Section 4 to produce the various data products.

Show mapping from data product type to category. i.e prompt contains images, catalogs, but not he same ones as DR/

UG catalogs can be federated with DR/PP catalogs.

These data product categories are defined in the SRD ((Ivezić & The LSST Science Collaboration 2018)) and have been a driver for DM (add more detail about why)

5.3. *Special programs*

Say something about data products from Special Programs. The special programs data products will be processed and stored as for all other data products. Maybe doesn't need to be a subsection

5.4. *Intermediate data products*

During processing, many intermediate data products are created. If is not feasible nor efficient to store them all. The DM system provides services to generate data products. Desscribe the generation of intermediate data products, in particular to generate intermediate flavours of coadds.

5.5. *Data Product Verification*

The quality of all Level 1 and Level 2 data products will be extensively assessed, both automatically as well as manually. Described in detail in PST (2020) Not sure this section is needed.

DP verification and QA will be carried out automatically following reprocessing runs and in advance of each data release Validation of the products is covered in PST (2020).

5.6. *Commissioning Data Products*

Describe here (possibly in a table) the exact data products delivered in commissioning and the science that they enable. This will be a high level description and can refer to the DP2 DPDD for exact details.

5.7. *Anticipated Data Release 1 Data Products*

The data products delivered during commissioning represent only an initial set of data products. The commissioning data set does not enable the production of .... (galaxy shape measurements ...) In addition to the commissioning data products described in § 5.6, we anticipate providing:

- `Object` catalog photo-z data products.

This list is neither definitive nor exhaustive; the exact list of data products to be provided as part of future LSST Data Releases will be determined closer to the time of release.

6. DATA MANAGEMENT TRANSITION TO OPERATIONS

Some of the ops proposal text ..

7. CHALLENGES

Remaining challenges perhaps ?

8. CONCLUSION

APPENDIX

REFERENCES

2020, LSST Data Management System Verification and Validation

Bauer, A. 2020, Overview of LSST Education and Public Outreach

Becla, J., Economou, F., Gelman, M., et al. 2016, Project Management Guide

Blum, R., & the Rubin Operations Team. 2020, Vera C. Rubin Observatory Data Policy

Dubois-Felsmann, G. P., Ivezic, Z., & Juric, M. 2018, LSST Data Product Categories

Gressler, W., DeVries, J., Hileman, E., et al. 2014, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9145, Ground-based and Airborne Telescopes V, ed. L. M. Stepp, R. Gilmozzi, & H. J. Hall, 1

Ivezić, Ž., & The LSST Science Collaboration. 2018, LSST Science Requirements Document

Jenness, T. 2020, LSST Data Management Software System

Jurić, M., Ciardi, D., & Dubois-Felsmann, G. 2017, LSST Science Platform Vision Document

Jurić, M., Kantor, J., Lim, K. T., et al. 2017, in ASP Conf. Ser., Vol. 512, Astronomical Data Analysis Software and Systems XXV, ed. N. P. F. Lorente, K. Shortridge, & R. Wayth, 279

Kahn, S. M., Kurita, N., Gilmore, K., et al. 2010, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7735, Ground-based and Airborne Instrumentation for Astronomy III, ed. I. S. McLean, S. K. Ramsay, & H. Takami, 0

Larman, C., & Basili, V. R. 2003, Computer, 36, 47

Lim, K.-T., Bosch, J., Dubois-Felsmann, G., et al. 2018, Data Management System Design

O'Mullane, W., Swinbank, J., Jurić, M., & DMLT. 2018, Data Management Organization and Management

Selvy, B. M., Claver, C., & Angeli, G. 2014, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9150, Modeling, Systems Engineering, and Project Management for Astronomy VI, ed. G. Z. Angeli & P. Dierickx, 0

A. ACRONYMS

Acronym	Description
CAM	CAMera
COVID	COrona VIrus Disease
DESC	Dark Energy Science Collaboration
DIA	Difference Image Analysis
DM	Data Management
DMLT	DM Leadership Team
DMTN	DM Technical Note
DP	Data Production
DP2	Data Preview 2
DPDD	Data Product Definition Document
DR	Data Release
DR1	Data Release 1
DRP	Data Release Production
EVMS	Earned Value Management System
FrDF	French Data Facility
GPL	GNU Public License
IN2P3	Institut National de Physique Nucléaire et de Physique des Particules
LDM	LSST Data Management (Document Handle)
LPM	LSST Project Management (Document Handle)
LSE	LSST Systems Engineering (Document Handle)
LSST	Legacy Survey of Space and Time (formerly Large Synoptic Survey Telescope)
NSF	National Science Foundation
PSTN	Project Science Technical Note
PVI	Processed Visit Image
QA	Quality Assurance
QC	Quality Control
RDO	Rubin Directors Office
RFC	Request For Comment
ROE	Royal Observatory Edinburgh
SLAC	SLAC National Accelerator Laboratory
SQuaRE	Science Quality and Reliability Engineering
SRD	LSST Science Requirements; LPM-17
T/CAM	Technical/Control (or Cost) Account Manager
UK	United Kingdom
US	United States
USDF	United States Data Facility
WBS	Work Breakdown Structure