Draft version November 12, 2025
Typeset using LaTeX **twocolumn** style in AASTeX7

Overview of the Vera C. Rubin Observatory Data Management

William O'Mullane [1], Leanne P. Guy [1], Yusra AlSayyad [2], Frossie Economou [3], Tim Jenness [3], Eric C. Bellm [4], Ian S. Sullivan [4], James F. Bosch [2], Gregory P. Dubois-Felsmann [5], Richard Dubois [6], Kian-Tat Lim [6], Fabio Hernandez [7], Mark G. Beckett [8], Mario Jurić [9], Jeffrey P. Kantor [3], Jacek Becla [6], Fritz Mueller [6], Stephen R. Pietrowicz [10], Colin T. Slater [4], and John D. Swinbank [11,2]

[1] *Vera C. Rubin Observatory, Avenida Juan Cisternas #1500, La Serena, Chile*
[2] *Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544, USA*
[3] *Vera C. Rubin Observatory Project Office, 950 N. Cherry Ave., Tucson, AZ 85719, USA*
[4] *University of Washington, Dept. of Astronomy, Box 351580, Seattle, WA 98195, USA*
[5] *Caltech/IPAC, California Institute of Technology, MS 100-22, Pasadena, CA 91125-2200, USA*
[6] *SLAC National Accelerator Laboratory, 2575 Sand Hill Rd., Menlo Park, CA 94025, USA*
[7] *CNRS/IN2P3, CC-IN2P3, 21 avenue Pierre de Coubertin, F-69627 Villeurbanne, France*
[8] *Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK*
[9] *Institute for Data-intensive Research in Astrophysics and Cosmology, University of Washington, 3910 15th Avenue NE, Seattle, WA 98195, USA*
[10] *NCSA, University of Illinois at Urbana-Champaign, 1205 W. Clark St., Urbana, IL 61801, USA*
[11] *ASTRON, Oude Hoogeveensedijk 4, 7991 PD, Dwingeloo, The Netherlands*

ABSTRACT

Vera C. Rubin Observatory Data Management (DM) subsystem is one of four construction subsystems. In operations we retain the notion of four departments of which one is DM. In this paper we describe DM as built as well as the fabric around DM which enabled its success. The goal of DM in construction was 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." That said we do outline the data products which will be produced by DM software as a part of the overall Rubin effort. We refer to detail oriented papers in many areas for the interested reader.

Keywords: Astrophysics - Instrumentation and Methods for Astrophysics — methods: data analysis — methods: miscellaneous

1. INTRODUCTION

Within the Vera C. Rubin Observatory (Ž. Ivezić et al. 2019) the Data Management (DM) team 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 Rubin Observatory Data Management System (M. Jurić et al. 2017; W. O'Mullane et al. 2022)

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. The 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	–
N Images per night	1000	–

1.2. *Technical Challenges*

The operational goal of Rubin Observatories Legacy Survey of Space and time is to produce an optical/near-IR survey of half the sky in ugrizy bands to r 27.5 (36 nJy) based on 825 visits over a 10-year period. It is a deep wide fast survey. Each Rubin image is around

8GB and we take more than one per minute or about 1000 per night. Add the ≈ 450 calibration exposures each day and it means about 20TB of data has to be shipped from Chile to SLAC on a daily basis. Alerts are to be produced in under 2 minutes with a goal of doing them in 1 minute which gives us a challenging transmission and prompt processing time window (see subsection 4.1).

Over the 10 year survey we estimate having 2.75M on sky images and at least 1M more calibrations totaling about 50PB. As we grow this must be reprocessed each year to produce the data releases (see subsection 4.6.

2. ORGANISATION OF DATA MANAGEMENT

The Organisation and management of DM constructionist covered in detail in W. O'Mullane et al. (2023). As shown in Figure 1 DM Management is aligned mainly along the Work 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 an DM will be verified and validated as part of System verification and validation (B. M. Selvy et al. 2014)

2.1. *Open development process*

From the outset DM was seen as a large scientific software project and the team evolved a unified development process (T. Jenness et al. 2018). Agile methodologies (C. Larman & V. R. 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.

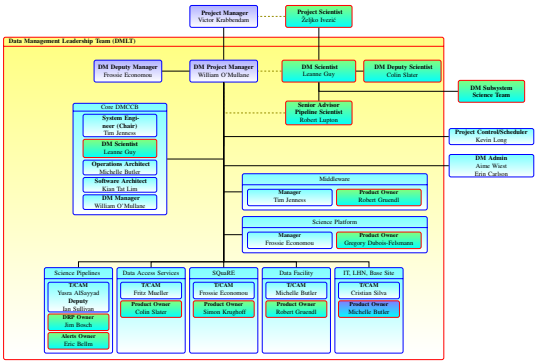

Figure 1. Org chart for Rubin DM from W. O'Mullane et al. (2023)

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 but some libraries that are particularly suited to more wider adoption use the BSD 3-clause license) 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 J. Becla et al. (2018); J. Kantor et al. (2016).

2.2. *Mode of work*

The DM team is distributed in several centers across the continental US as well as Chile, France and the UK. 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 helped 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, before the pandemic and morphed to one physical and two or three virtual meetings a year. The Managers have a standup meeting on Thursdays to work out any blockages or an-

ticipated 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). As we approached operations we also introduced the Rubin Technical Notes (RTN series).

2.3. *Relationship to other subsystems*

In construction the subsystems started quite distinctly leading to a certain amount of *siloing*. Early on this is good to allow the project to quickly start on many fronts but later, for integration, more communication is required to make sure the parts match up. Hence early in the project DM had little interaction with other subsystems and teams apart from System Engineering, in the end it is much more involved. The System Engineering team interacted with all parts of the project especially in the area of requirements engineering. DM used the system engineering tools such as Rose and later magic draw which supported the Model Based System Engineering approach C. F. Claver et al. (2014) DM built on this relationship and created some tools to aid the verification process around Jira.

Interactions with the other subsystems are governed by change controlled Interface Control Documents (ICD), though in many cases these were more like requirements documents we kept them updated through construction so the reflect the as built system going into operations. The main interface for DM was to the LSST camera (S. M. Kahn et al. 2010) and LATISS (P. Ingraham et al. 2020) to capture images and spectra. Image capture was originally an over designed parallel system where DM directly called the camera interface and reconstructed the images. It was felt this was error prone and would lead to discrepancies between DM and Camera. A simpler image interface was proposed (K.-T. Lim 2022) whereby Camera writes the image file including the header and DM then pick it up for transfer. The header is provided by a DM service which listens to several telescope and site topics to gather information.

The interface to telescope and site was defined to be through the System Abstraction Layer (SAL) and large remains. This message bus system allows for commands to be sent to components as well as components to listen to messages from other components. In addition to picking up header information DM sends near realtime image metric information through SAL which is then displayed in the LSST Operations Visualisation Environment(LOVE). When DM performs actions, such as a header being ready, this is also broadcast through SAL.

Though not in the original design DM supplied the underlying Engineering Facilities Database (EFD) service to telescope and site starting in 2018. TODO: Cite the EFD SPIE paper

DM also have interactions with Education and Public Outreach (EPO). We provide EPO with some of the EFD data and up to 10% of the image data for public use. In addition DM produce a set of color HiPS **?** maps for EPO to use in their interactive browser - these use a different color map to the science HiPS maps used.

Through the Chile DevOps team telescope and site is supported by providing computing infrastructure and fibre optic cabling in the observatory. In a breaking down of silos telescope and site software uses Dm like infrastructure with most components deployed via kubernetes and phalanx. This includes using conda for dependency management, github for code and lsst.io for documentation. A few summit systems, including camera machines, use only puppet. A few National Instruments base systems can not be automated and need manual upgrades which are generally not handled by DM.

We are commanded and listen to the telescope and site software (S. J. Thomas et al. 2022)

2.4. *Data Management transition to operations*

DM aimed for a smooth transition to operations maintaining many of the team members though some DM team members move to System Performance. The organisation structure was rationalized to match the cyber infrastructure layout in operations and is depicted in Figure 2. The logical make up of this following out data taking to data serving approach is depicted in Figure 5. Leadership remains very similar assisting the transition.

DM transitioned parts of the system over many years, this was a reality but partly it was in response to the availability of operational funding requiring us to be more clear about this. For example once the Science Platform on the summit became a daily requirement it was considered operational and changes were rigorously controlled this was delivered in 2021 (F. Economou et al. 2021). The EFD was officially delivered in 2023 though it has actually been in operations supporting summit activities since at least 2019. The change from NCSA to a DOE Data facility in 2020 lead to the creation of the the Interim Data Facility and the deployment of user facing operational services on Google(W. O'Mullane et al. 2024). This allowed DM to demonstrate the ability to

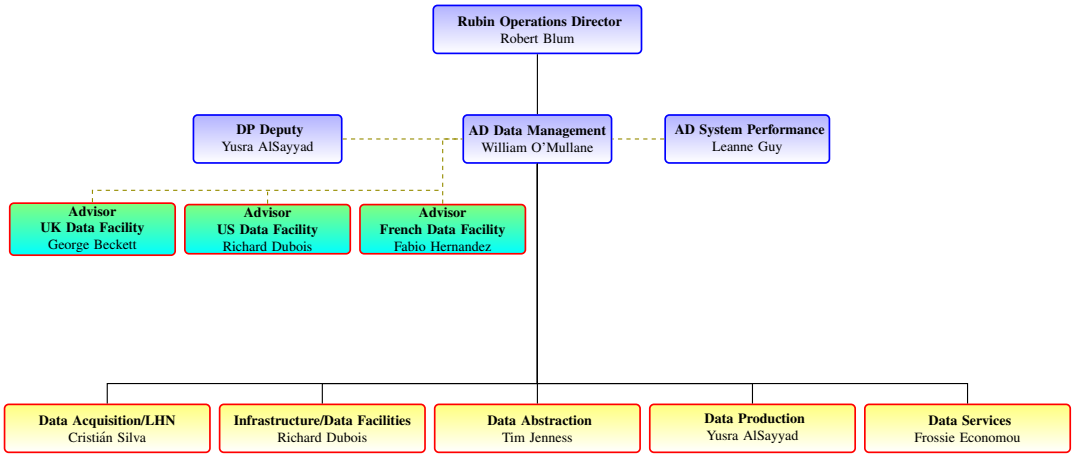

Figure 2. The operations organisation follows the cyber infrastructure layout of DM more than the WBS structure of construction (from W. O'Mullane (2025)). It is somewhat simpler than Figure 1.

operate all of data production and data serving which exercised data abstraction and infrastructure through three data previews with simulated data. This was not all of DM operations but a significant fraction the other main part, data acquisition, was being exercised since 2019 with regular Auxiliary Telescope runs.

Hence over five year DM gradually transitioned to operations while maintaining our ability to develop and improve our systems as we intend to do throughout operations.

3. ARCHITECTURE, DATA TRANSMISSION AND ACCESS

DM spans multiple locations with processing occurring at the USDF (SLAC), FrDF (IN2P3) and UK (IRIS). The system vision has been fairly consistently to deliver science ready data products to the Rubin community as depicted in Figure 3.

The organisation and management of DM is covered in section 2

The DM system architecture was laid out in K.-T. Lim et al. (2020) from which we reproduce Figure 4

As shown in Figure 3 there are several kinds of Rubin data - mostly they are accessed via the science platform or other services which are described in subsection 4.7.

Data production subsection 4.6 is responsible for all of the pipelines and their execution.

Of course all these services and pipelines must run on hardware which is typically at a data facility. The data facilities are covered in subsection 4.8

3.1. *Alerts and Brokers*

Alerts are product of DM operations and briefly covered in section 5. The software producing alerts, know and the Alerts Pipeline (AP), is discussed in TODO: REFER TO ALERTS PIPELINES SECTION TODO: Leanne: you said you might have a go a this Mention community alert brokers.

4. DM SOFTWARE SYSTEMS

The DM products are not data as many may think, rather the products are software and services to produce those products. The management and organisation of DM change slightly for operations (see subsection 2.4) but many of the same people have similar operations roles giving a good continuity. Going into operations we assessed the way DM works and reconceptualized the organisation around the data flow and cyber infrastructure.

The high level list of DM products is given in Figure 6, 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 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. *Data Acquisition*

4.2. *Data Abstraction*

The Data Abstraction is responsible for providing standardized interfaces to data and metadata such that

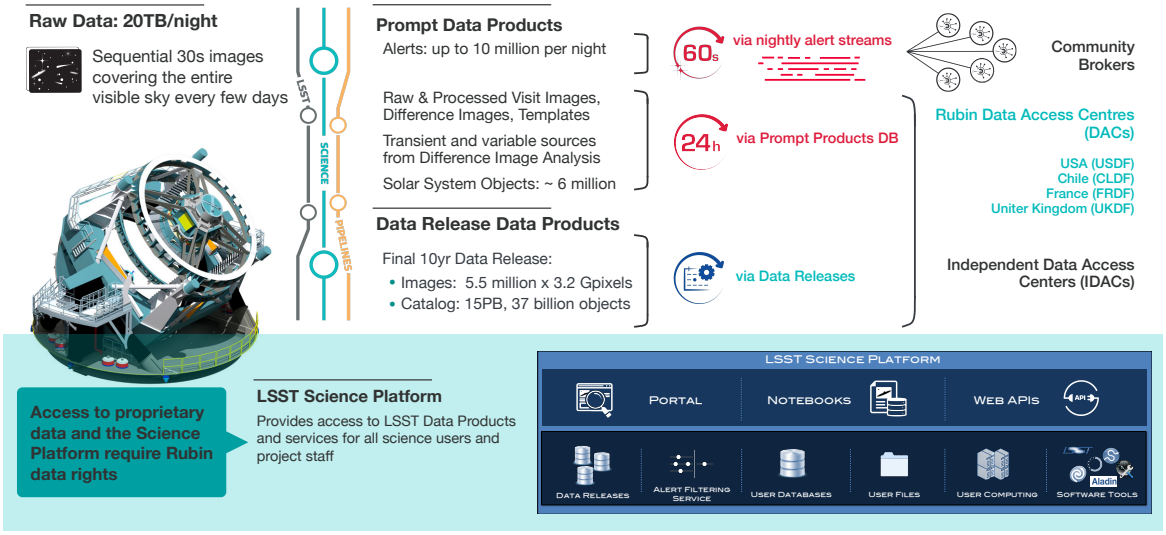

Figure 3. Overview of data management from the telescope to the user.

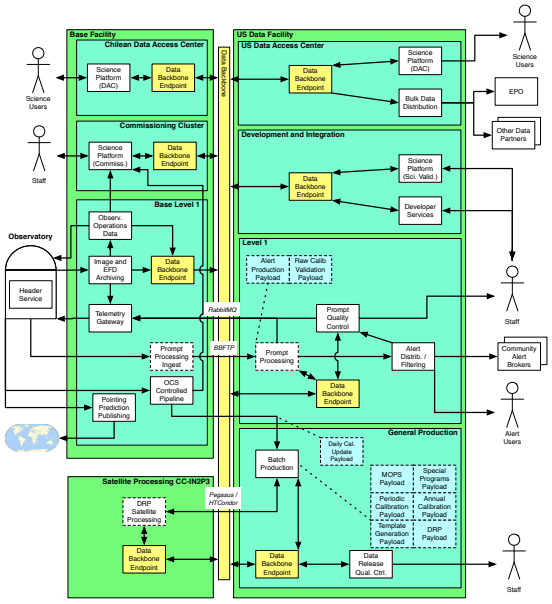

Figure 4. Rubin DM architecture diagram K.-T. Lim et al. (2020)

the science users and pipeline developers can focus on the algorithms and science results.

4.3. *Data Engineering*

The Data Engineering team:

- Validates file FITS headers and provides tooling for ensuring correct values are stored in the headers even if the file was originally written incorrectly.
- Provides standardized metadata translation mechanisms such that downstream users can always ask for information from an observation regardless of the instrument or instrument-specific FITS header conventions.
- Provides tooling for specifying schemas used for data release products in a machine-readable form (W. O'Mullane & C. Slater 2020).
- Follows and contributes to evolving IVOA standards.

4.4. *Pipeline Middleware*

From the very beginning of the project it was decided that algorithm code should always work on in-memory representations of datasets and should not know where data come from, what form it was stored on disk, or where data will be written to or how it will be written. The Data Butler was developed to meet these requirements (T. Jenness 2024; T. Jenness et al. 2022).

4.5. *Build Engineering*

- Use Jenkins to make pipelines releases and to support continuous integration.
- Use EUPS and Docker for distribution.

4.6. *Data Production*

Data Production is underpinned by the fast and robust LSST Science Pipelines (Rubin Observatory Science Pipelines Developers 2025; J. Bosch et al. 2019), the image processing software written to convert the raw pixels from the Rubin observatory into science-ready data products for astronomers. It takes the raw images as input, and calibrates away the effects of the instrument and atmosphere to produce catalogs and images. Many science analyses can be done with the catalogs

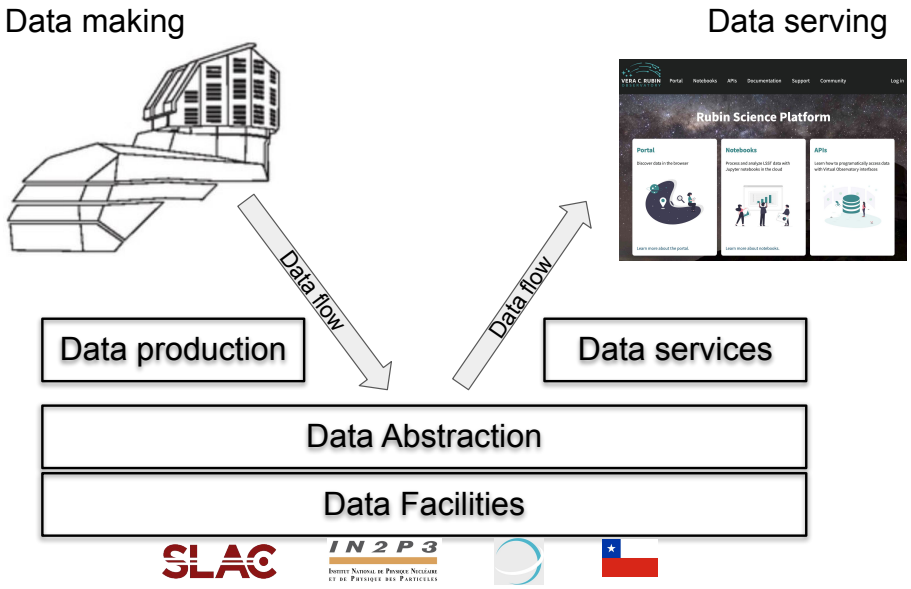

Figure 5. Rubin DM organisation in terms of data taking to data serving supported by cyber infrastructure.

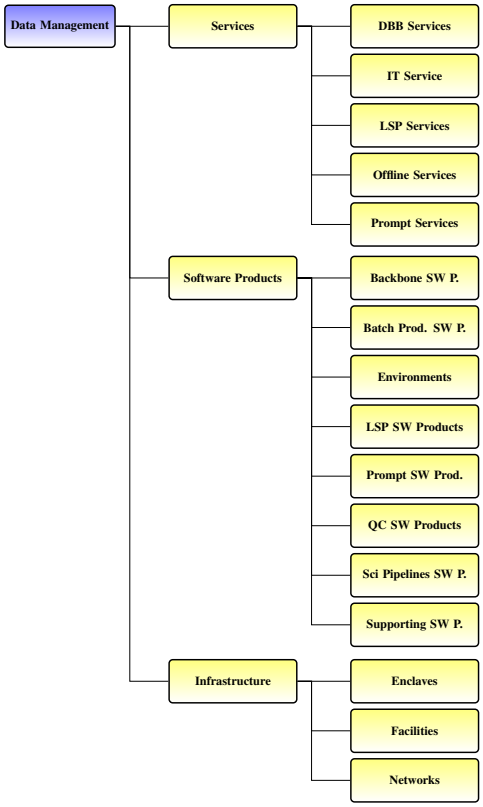

Figure 6. Rubin DM product tree

alone. Still, as new image processing algorithms are developed over the next decade, we expect the output calibrated coadds, difference images and processed visit images to be used by scientists running specialized detection algorithms during the survey.

The LSST Science Pipelines deliver data products fast and slow. The prompt data products are delivered via the nightly alert stream. These data products support science that requires rapid follow-up. The slower annual data release processing produces calibrated images and catalogs, including lightcurves, to support static sky science and statistical studies of variability.

These pipelines incorporate algorithms for tasks such as detrending, image subtraction, deblending, object characterization, and sky background estimation, among others. When research and development on the pipelines first began in 2004, there was code that could accomplish some of these routines (AstroPy, PyRAF), but none were as robust and fast as needed for LSST. With 3.2 gigapixels of data rolling in every 30 seconds, the data volume grows very quickly. Fast and robust algorithms are needed to process this data efficiently.

The pipelines achieve their speed through Python3-wrapped C++ and are versatile enough for any ground-based optical or IR telescope. However, they require well-sampled PSFs, making them unsuitable for space-based imaging.

The LSST Science Pipelines will continue to evolve throughout LSST's 10-year survey. A portion of LSST's operating budget will be spent on maintaining state-of-the-art algorithms. The state of the art has changed

significantly over the last ten years, and there's no reason to believe it will not change over the next decade.

The current algorithms reflect the hard-earned lessons from precursor surveys such as the Dark Energy Survey and Pan-STARRS. These include, for example:TODO: Yusra check these refs are what you wanted

- the PSF modeling algorithm, PIFF (citation)
- the astrometric calibration algorithms GBDES. (G. M. Bernstein et al. 2017)
- The photometric calibration algorithm, FGCM (D. L. Burke et al. 2018)
- The artifact rejection algorithm during coaddition. (citation)
- pattern continuity algorithm for matching amp-to-amp gain offsets (citation?)

Formal Agile development practices were adopted in 2014 when we received funding to start construction. At the time, we had minimal-viable algorithm pipelines used in both internal data challenges to process SDSS Stripe 82 data (J. Kantor 2010; M. Juric 2012; M. Juric et al. 2013), and they were also selected as the data release pipelines for the Hyper SuprimeCam Strategic Survey Program (J. Bosch et al. 2018). Feedback from the scientific community, particularly through four public data releases of the Hyper Suprime Cam (HSC) data, has been crucial in refining our algorithms.

We combine unit tests, continuous integration tests, and regression tests. During construction, Jenkins runs continuous integration tests nightly on small subsets of precursor data, including simulated LSST data and public HSC data. Before merging with the main branch, developers test their ticket branches on these CI tests.

The science pipelines are run in prompt and data release production, utilizing the DM Middleware task framework (Section 4.2). This abstraction layer significantly enhances the portability of science pipelines. The Butler acts as a data abstraction layer, removing the need for direct I/O operations or knowledge of the storage backend by the pipelines. Data releases have been successfully executed using the pipelines on Google Cloud and on-premise hardware, managed by workflow systems such as HTCondor or PanDA. The primary startup cost involves ingesting your dataset into the Butler.

All algorithms are implemented as subclasses of the parent PipelineTask, which specifies their inputs and outputs. This structure enables the middleware to construct a directed acyclic graph of all processing tasks required for a specific data product. These tasks are the fundamental building blocks of the pipelines. The pipelines themselves consist of these tasks, each utilized in various ways across different processes. For instance, the data release and other production pipelines include the same subtractImages task.

Initially, the Science Pipelines were designed to run exclusively on CPUs, reflecting the hardware budgeting at the start of construction. Our processes are highly parallelizable, and we anticipate utilizing tens of thousands of cores during data release processing, with each core dedicated to a specific region of the sky or a particular observation. Given the available RAM per core, optimal sizing of sky patches could lead to full CPU utilization. Given advancements in image processing, we are also considering the potential integration of GPUs.

Documentation and installation instructions can be found at pipelines.lsst.io.

4.7. *Data Services*

text here

4.8. *Data Facilities*

As noted in section 3, data processing will occur at three data facilities — in USA, France, and UK. In particular, preparation of the (typically, annual) Data Releases will be distributed across these three facilities using specialised software tools and techniques for distributed data management and remote job submission adopted from the high-energy physics community, with DM providing the required interfaces to the Science Pipeline.

In this arrangement, the USDF will coordinate each processing *Campaign* and be the primary curation site, holding a copy of all raw, intermediate, and science-ready products from each production run of the Science Pipeline. The USDF will also be solely responsible for Prompt Processing.

4.8.1. *US Data Facility*

4.8.2. *French Data Facility*

The computing centre of France's National institute of nuclear and particle physics (IN2P3)[12] hosts and operates Rubin's French Data Facility (FrDF)[13]. This computing and storage infrastructure is sized to store a full copy of the raw images as well as to contribute 40% of the image processing capacity required to produce the Data Releases, for the duration of the observatory's operations phase.

[12] https://cc.in2p3.fr
[13] https://doc.lsst.eu

A compute element exposes the site's batch farm to Rubin's central campaign management system and a Butler-compatible storage element (see subsection 4.2) stores input data as well as locally-produced data products. At the end of each processing campaign, final products are replicated to the US Data Facility where they are combined for composing the Data Release.

FrDF builds and packages the LSST Science Pipelines for distribution via a software content distribution based on CERN's CernVM File System[14]. This distribution mechanism, which all the Rubin data facilities subscribe to, ensures that they all use an identical copy of the pipelines for the purposes of producing the Data Releases.

In addition, the French Data Facility contributes to perform realistic test campaigns of Rubin's distributed system being developed to prepare the Data Releases, including the development of the inter-facility data replication system. Evaluation instances of the Rubin Science Platform and the catalog database have been locally deployed continuously since several years. The facility also hosts Fink (A. Möller et al. 2020), one of the Rubin community alert brokers.

4.8.3. *UK Data Facility*

UK interest in the Vera C. Rubin Observatory is coordinated by the LSST:UK Consortium, which has 36 partners representing all major UK astronomy research groups.

Via the Rubin In-kind Contribution program, LSST:UK has proposed — among other things — to provide computing resources and associated staff time to undertake 25% of the computing associated with the preparation of each Data Release.

The infrastructure (the UK Data Facility) for this and other significant in-kind contributions has been secured from the UK IRIS program[15] on a mix of grid, high-performance and research cloud facilities.

In particular, it is proposed that Data Release Processing will occur on grid-computing services at Lancaster University and Rutherford Appleton Laboratories (RAL). Staff at Lancaster and RAL are directly involved in the development of the distributed DRP approach with particular contributions to data distribution and progress tracking, job handling, and infrastructure health monitoring.

LSST:UK has also proposed to operate a full Independent Data Access Center (IDAC), with capacity to serve the two most recent Data Releases to 20% of the anticipated Rubin international community via the Rubin Science Platform.

The UK IDAC is an integral part of the UK Data Facility, mostly hosted in on-premises cloud resources at the University of Edinburgh, though with some ancillary services provided by RAL. At the time of writing, LSST:UK has been running a prototype IDAC for more than two years, hosting precursor and ancillary astronomy surveys for 20 or so early adopters.

Other contributions that are provided by the UK Data Facility include a Rubin Community Broker, called Lasair, and an HPC-based instance of the Science Pipeline for the production of specific User-generated Products that support the fusion of LSST with compatible near-infrared surveys and the crossmatch of LSST object catalogues with contemporary surveys.

5. DATA PRODUCTS

Rubin Observatory's LSST Science Pipelines (§ 4) will produce the *science-ready data products*. These data products have been carefully designed to enable the vast majority of LSST science without the need to access the raw pixels, nor for users to reprocess the data. There will however be some science cases where pixel access or a reprocessing of the data is warranted are, such as estimating and subtracting a different background (LSB science), reprocessing a small fraction of images to develop the systematics budget for weak lensing studies (Dark Energy science), or injecting fake objects into images and reprocessing them to develop models for artifact rejection. In all such cases involving image reprocessing, we anticipate that users will start from images that have been corrected for instrumental effects and photometrically and astrometrically calibrated.

The Data Products Definition Document, (M. Jurić et al. 2023) was used to describe the data products produced by the LLSST and guide the development of the Data Management System.

In this section we provide a high-level overview of the LSST science-ready data products. A detailed description of the LSST data products and their scientific performance on the early LSST commissioning data is given in (L. P. Guy 2019).

5.1. *Types of Data Product*

LSST produces several types of data products.

`Images` — processed 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)

[14] https://sw.lsst.eu

[15] www.iris.ac.uk — last accessed May 24th, 2024.

Rubin images are rich data products, which, in addition to storing the image pixel data also contain the PSF model, WCS and mask plane, ... what else

Include a description of cutout images and how they will be accessed

What is the maximum size of a cutout, how many at a time?

Image data products also includes calibration frames (darks, flats, biases, fringe, etc.)

coadds – We reiterate that not all coadds will be kept and served to the public

template coadds RGB color images derived from coadds

All calibration frames (darks, flats, biases, fringe, etc.) will be preserved and made available. Provide the full list of calibration images and the data products that come out of cp_pipe.

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

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

Object ‘ Source

ForcedSource

ShearObject

`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.

In addition to the alerts detected on DIASources above the nominal detection threshold of 5σ, we also measure and store a small sample of DIASources detected the nominal 5σ threshold. There are several drivers for these *sub-threshold alerts*, for example to enable monitoring of difference image analysis quality or to assess the danger posed by a potentially hazardous asteroid. A set of criteria, described in (E. Bellm et al. 2023) was defined based on key science cases.

`Calibration Data Products` —

`Survey Property Maps` —

Several types of survey property maps will be generated and served to users. The properties are typically the mean or total values determined from the images input to generate the deep coadd. The types of maps will include the total exposure time; the point-source 5-sigma AB magnitude limit; the weighted mean of the PSF moments; the weighted mean of the sky background and sky noise; and the average effect of differential chromatic refraction (DCR) in the right ascension and declination directions, and in the PSF moments. Property maps based on statistics measured on deep coadds might also be generated.

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.

5.3. *Other categories of data products*

5.3.1. *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 G. Dubois-Felsmann et al. (2018)

In operations Data Production will use 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 (Ž. Ivezić & The LSST Science Collaboration 2018) and have been a driver for DM (add more detail about why)

5.4. *Special programs data products*

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

5.5. *Custom 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. Describe the generation of custom data products, in particular to generate flavours of coadds.

6. CHALLENGES

Remaining challenges perhaps ?

7. CONCLUSION

ACKNOWLEDGMENTS

This material is based upon work supported in part by the National Science Foundation through Cooperative Agreement AST-1258333 and Cooperative Support Agreement AST-1202910 managed by the Association of Universities for Research in Astronomy (AURA), and the Department of Energy under Contract No. DE-AC02-76SF00515 with the SLAC National Accelerator Laboratory managed by Stanford University. Additional Rubin Observatory funding comes from private donations, grants to universities, and in-kind support from LSSTC Institutional Members.

Facilities: Rubin:Simonyi (LSSTCam), Rubin:1.2m (LATISS)

Software: Rubin Science Platform (G. Dubois-Felsmann et al. 2019), LSST Science Pipelines (Rubin Observatory Science Pipelines Developers 2025), Qserv (D. L. Wang et al. 2011)

REFERENCES

Becla, J., Economou, F., Gelman, M., et al. 2018,, Data Management Technical Note DMTN-020, NSF-DOE Vera C. Rubin Observatory. https://dmtn-020.lsst.io/

Bellm, E., Graham, M., Guy, L., & The DM System Science Team. 2023,, Data Management Technical Note DMTN-228, NSF-DOE Vera C. Rubin Observatory. https://dmtn-228.lsst.io/

Table 2. Summary of LSST Data Products and the cadences on which they will be released

LSST Data Product	Prompt			Data Release
	60/120 sec	$\leq$ 24 hrs	$\geq$ 80 hrs	$\approx$ Annually
Images	Cut-outs in Alerts	–	Raw Images Processed Visit Images Difference Images Image Differencing Templates	Raw exposures Processed Visit Images Calibration frames Deep coadds Template coadds RGB Images
Catalogs	Prompt Catalog data in Alerts	Prompt Catalogs	—	Data Release Catalogs
Alerts	Alerts from DIASources	—	—	—
Maps	—	—	—	Survey property maps
Spectra	—	—	—	Auxiliary Telescope spectra

Bernstein, G. M., Armstrong, R., Plazas, A. A., et al. 2017, PASP, 129, 074503, doi: 10.1088/1538-3873/aa6c55

Bosch, J., Armstrong, R., Bickerton, S., et al. 2018, PASJ, 70, S5, doi: 10.1093/pasj/psx080

Bosch, J., AlSayyad, Y., Armstrong, R., et al. 2019, in Astronomical Society of the Pacific Conference Series, Vol. 523, Astronomical Data Analysis Software and Systems XXVII, ed. P. J. Teuben, M. W. Pound, B. A. Thomas, & E. M. Warner, 521, doi: 10.48550/arXiv.1812.03248

Burke, D. L., Rykoff, E. S., Allam, S., et al. 2018, AJ, 155, 41, doi: 10.3847/1538-3881/aa9f22

Claver, C. F., Selvy, B. M., Angeli, G., et al. 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, 91500M, doi: 10.1117/12.2056781

Dubois-Felsmann, G., Economou, F., Lim, K.-T., et al. 2019,, Data Management Controlled Document LDM-542, NSF-DOE Vera C. Rubin Observatory. https://ldm-542.lsst.io/

Dubois-Felsmann, G., Ivezic, Z., & Juric, M. 2018,, Project Controlled Document LPM-231, NSF-DOE Vera C. Rubin Observatory. https://lpm-231.lsst.io/

Economou, F., Thornton, A., Banek, C., Allbery, R., & Krughoff, S. 2021,, Technical Note RTN-019, NSF-DOE Vera C. Rubin Observatory. https://rtn-019.lsst.io/

Guy, L. P. 2019,, Project Science Technical Note PSTN-024, NSF-DOE Vera C. Rubin Observatory. https://pstn-024.lsst.io/

Ingraham, P., Clements, A. W., Ribeiro, T., et al. 2020, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 11452, Software and Cyberinfrastructure for Astronomy VI, ed. J. C. Guzman & J. Ibsen, 114520U, doi: 10.1117/12.2561112

Ivezić, Ž., & The LSST Science Collaboration. 2018,, Project Controlled Document LPM-17, NSF-DOE Vera C. Rubin Observatory. https://ls.st/LPM-17

Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111, doi: 10.3847/1538-4357/ab042c

Jenness, T. 2024,, Data Management Technical Note DMTN-288, NSF-DOE Vera C. Rubin Observatory. https://dmtn-288.lsst.io/

Jenness, T., Economou, F., Findeisen, K., et al. 2018, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 10707, Software and Cyberinfrastructure for Astronomy V, ed. J. C. Guzman & J. Ibsen, 1070709, doi: 10.1117/12.2312157

Jenness, T., Bosch, J. F., Salnikov, A., et al. 2022, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 12189, Software and Cyberinfrastructure for Astronomy VII, 1218911, doi: 10.1117/12.2629569

Juric, M. 2012,, Data Management Technical Note DMTN-034, NSF-DOE Vera C. Rubin Observatory. https://dmtn-034.lsst.io/

Juric, M., Becker, A., Shaw, R., Krughoff, K. S., & Kantor, J. 2013,, Data Management Technical Note DMTN-035, NSF-DOE Vera C. Rubin Observatory. https://dmtn-035.lsst.io/

Jurić, M., Kantor, J., Lim, K. T., et al. 2017, in Astronomical Society of the Pacific Conference Series, Vol. 512, Astronomical Data Analysis Software and Systems XXV, ed. N. P. F. Lorente, K. Shortridge, & R. Wayth, 279, doi: 10.48550/arXiv.1512.07914

Jurić, M., Axelrod, T. S., Becker, A. C., et al. 2023,, Systems Engineering Controlled Document LSE-163, NSF-DOE Vera C. Rubin Observatory, doi: 10.71929/rubin/2587118

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, 77350J, doi: 10.1117/12.857920

Kantor, J. 2010, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7740, Software and Cyberinfrastructure for Astronomy, ed. N. M. Radziwill & A. Bridger, 77401O, doi: 10.1117/12.857253

Kantor, J., Long, K., Becla, J., et al. 2016, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9911, Modeling, Systems Engineering, and Project Management for Astronomy VI, ed. G. Z. Angeli & P. Dierickx, 99110N, doi: 10.1117/12.2233380

Larman, C., & Basili, V. R. 2003, Computer, 36, 47, doi: 10.1109/MC.2003.1204375

Lim, K.-T. 2022,, Data Management Technical Note DMTN-143, NSF-DOE Vera C. Rubin Observatory. https://dmtn-143.lsst.io/

Lim, K.-T., Bosch, J., Dubois-Felsmann, G., et al. 2020,, Data Management Controlled Document LDM-148, NSF-DOE Vera C. Rubin Observatory, doi: 10.71929/rubin/2587850

Möller, A., Peloton, J., Ishida, E. E. O., et al. 2020, Monthly Notices of the Royal Astronomical Society, 501, 3272, doi: 10.1093/mnras/staa3602

O'Mullane, W. 2025,, Technical Note RTN-046, NSF-DOE Vera C. Rubin Observatory. https://rtn-046.lsst.io/

O'Mullane, W., & Slater, C. 2020,, Data Management Technical Note DMTN-153, NSF-DOE Vera C. Rubin Observatory. https://dmtn-153.lsst.io/

O'Mullane, W., Swinbank, J., Juric, M., Guy, L., & DMLT. 2023,, Data Management Controlled Document LDM-294, NSF-DOE Vera C. Rubin Observatory. https://ldm-294.lsst.io/

O'Mullane, W., Economou, F., Lim, K.-T., et al. 2022, arXiv e-prints, arXiv:2211.13611, doi: 10.48550/arXiv.2211.13611

O'Mullane, W., Economou, F., Huang, F., et al. 2024, in Astronomical Society of the Pacific Conference Series, Vol. 535, Astromical Data Analysis Software and Systems XXXI, ed. B. V. Hugo, R. Van Rooyen, & O. M. Smirnov, 227, doi: 10.48550/arXiv.2111.15030

Rubin Observatory Science Pipelines Developers. 2025,, Project Science Technical Note PSTN-019, NSF-DOE Vera C. Rubin Observatory, doi: 10.71929/rubin/2570545

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, 91500N, doi: 10.1117/12.2056773

Thomas, S. J., Barr, J., Callahan, S., et al. 2022, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 12182, Ground-based and Airborne Telescopes IX, ed. H. K. Marshall, J. Spyromilio, & T. Usuda, 121820W, doi: 10.1117/12.2630226

Wang, D. L., Monkewitz, S. M., Lim, K.-T., & Becla, J. 2011, in State of the Practice Reports, SC '11 (New York, NY, USA: ACM), 12:1–12:11, doi: 10.1145/2063348.2063364