

**Centre de Calcul**  
de l'Institut National de Physique Nucléaire  
et de Physique des Particules

# Rubin Data Preview 0.2 at IN2P3

F. Hernandez, Q. Le Boulc'h

[Journées LSST-France](#), Annecy, May 16-18 2022

doc.lsst.eu

# Rubin data previews


- Datasets and analysis tools made available to a limited number of data right holders to begin early preparations for science with the LSST
- Useful also as integration tests of the [Rubin Science Platform](#) and the [LSST science pipelines](#)
- More details in [this Community post](#)



# Rubin Science Platform

### Portal

Discover data in the browser



Learn more about the portal.

### Notebooks


Process and analyze LSST data with Jupyter notebooks in the cloud



Learn more about notebooks.

### APIs

Learn how to programatically access data with Virtual Observatory interfaces



### TAP Searches

1. Select TAP Service ? Using LSST RSP <https://data.lsst.cloud/api/tap> - Replace...

2. Select Query Type ?  Single Table (UI assisted)  Edit ADQL (advanced)

3. Select Table ? Table Collection (Schema): **dp01\_dc2\_catalogs**  
 Data Preview 0.1 includes five tables based on the DESC's Data Challenge 2 simulation of 300 square degrees of the wide-fast-deep LSST su...  
 Table: **dp01\_dc2\_catalogs.object**  
 The object table from the DESC DC2 simulated sky survey as described in arXiv:2101.04855. Includes astrometric and photometric parameters...

4. Enter Constraints ? 34 of 137 columns selected [Reset Column Selections & Constraints](#)

**Spatial** ? *empty entry*

Longitude Column:  🔍  
 Latitude Column:  🔍

Shape Type:  ⬇

Coordinates or Object Name:  🚫 Try NED then Simbad ⬇  
*Examples: '62, -37' '60.4 -35.1' '4h11m59s -32d51m59s equ j2000' '239.2 -47.6 gal' 'NGC 1532' (NB: DC2 is a simulated sky, so names are not useful)*

Radius:   ⬇  
 Valid range between: 1" and 360000"

**Temporal** ?

#### Output Column Selection and Constraints

<input type="checkbox"/>	column_name <i>char</i>	constraints <i>char</i>	unit <i>char</i>	ucd <i>char</i>	description <i>char</i>	datatype <i>char</i>	arraysize <i>char</i>	utype <i>char</i>	xtype <i>char</i>	priority <i>int</i>
<input checked="" type="checkbox"/>	objectId			meta.id;src	Unique id.	long				
<input checked="" type="checkbox"/>	parentObjectId				Id of the parent object this object has been d	long				
<input checked="" type="checkbox"/>	ra		deg	pos.eq.ra	RA-coordinate of the center of the object for	double				
<input checked="" type="checkbox"/>	dec		deg	pos.eq.dec	Decl-coordinate of the center of the object fo	double				
<input checked="" type="checkbox"/>	blendedness				measure of how flux is affected by neighbors	double				
<input checked="" type="checkbox"/>	clean				True if the source has no flagged pixels and is	boolean				
<input checked="" type="checkbox"/>	cModelFlux_flag_g				Flag for issues with cModelFlux_flag_<band>	boolean				
<input checked="" type="checkbox"/>	cModelFlux_flag_i				Flag for issues with cModelFlux_flag_<band>	boolean				
<input checked="" type="checkbox"/>	cModelFlux_flag_r				Flag for issues with cModelFlux_flag_<band>	boolean				
<input checked="" type="checkbox"/>	cModelFlux_flag_u				Flag for issues with cModelFlux_flag_<band>	boolean				
<input checked="" type="checkbox"/>	cModelFlux_flag_y				Flag for issues with cModelFlux_flag_<band>	boolean				
<input checked="" type="checkbox"/>	cModelFlux_flag_z				Flag for issues with cModelFlux_flag_<band>	boolean				
<input type="checkbox"/>	cModelFlux_g				composite model (CModel) flux in _<band>	double				
<input type="checkbox"/>	cModelFlux_i				composite model (CModel) flux in _<band>	double				
<input type="checkbox"/>	cModelFlux_r				composite model (CModel) flux in _<band>	double				
<input type="checkbox"/>	cModelFlux_u				composite model (CModel) flux in _<band>	double				
<input type="checkbox"/>	cModelFlux_y				composite model (CModel) flux in _<band>	double				
<input type="checkbox"/>	cModelFlux_z				composite model (CModel) flux in _<band>	double				
<input type="checkbox"/>	cModelFluxErr_g				Error value for cModel flux in _<band>	double				
<input type="checkbox"/>	cModelFluxErr_i				Error value for cModel flux in _<band>	double				
<input type="checkbox"/>	cModelFluxErr_r				Error value for cModel flux in _<band>	double				
<input type="checkbox"/>	cModelFluxErr_u				Error value for cModel flux in _<band>	double				
<input type="checkbox"/>	cModelFluxErr_y				Error value for cModel flux in _<band>	double				

Filter files by name

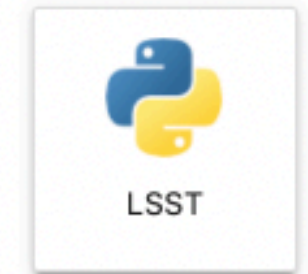
/ notebooks / tutorial-notebooks /

Name	Last Modified
data	3 minutes ago
01_Intro_to_DP0_Notebooks.ipynb	3 minutes ago
02_Intermediate_TAP_Query.ipynb	3 minutes ago
03_Image_Display_and_Manipulation.ipynb	3 minutes ago
03b_Image_Display_with_Firefly.ipynb	3 minutes ago
04_Intro_to_Butler.ipynb	3 minutes ago
05_Intro_to_Source_Detection.ipynb	3 minutes ago
06_Comparing_Object_and_Truth_Tables.ipynb	3 minutes ago
08a_Interactive_Image_Visualization.ipynb	3 minutes ago
08b_Interactive_Catalog_Visualization.ipynb	3 minutes ago
09_Single_Star_Lightcurve_with_Butler.ipynb	3 minutes ago
README.md	3 minutes ago

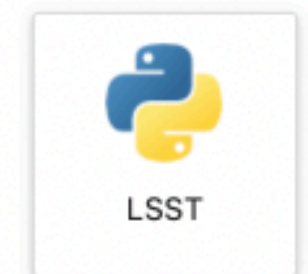
Launcher

notebooks/tutorial-notebooks

Notebook



Console



Other

  
Terminal

  
Text File

  
Markdown File

  
Python File

  
Show Contextual Help

# Rubin data release scenario



<b>Rubin Baseline Data Release Scenario</b>	Jun 2021	Jun 2022	Mar 2024 - Jun 2024	Jul 2024 - Dec 2024	Apr 2025 - Jul 2025	Apr 2026 - Jul 2026	Apr 2027 - Jul 2027	Apr 2028 - Jul 2028
	<b>DP0.1</b>	<b>DP0.2</b>	<b>DP1</b>	<b>DP2</b>	<b>DR1</b>	<b>DR2</b>	<b>DR3</b>	<b>DR4</b>
<b>Data Product</b>	DC2 Simulated Sky Survey	Reprocessed DC2 Survey	ComCam On-Sky Data	LSSTCam On-Sky Data	LSST First 6 Months Data	LSST Year 1 Data	LSST Year 2 Data	LSST Year 3 Data
Raw images	✓	✓	✓	✓	✓	✓	✓	✓
DRP Processed Visit Images and Visit Catalogs	✓	✓	✓	✓	✓	✓	✓	✓
DRP Coadded Images	✓	✓	✓	✓	✓	✓	✓	✓
DRP Object and ForcedSource Catalogs	✓	✓	✓	✓	✓	✓	✓	✓
DRP Difference Images and DIASources	☐	✓	✓	✓	✓	✓	✓	✓
DRP ForcedSource Catalogs including DIA outputs	☐	✓	✓	✓	✓	✓	✓	✓
PP Processed Visit Images	☐	☐	✓	✓	✓	✓	✓	✓
PP Difference Images	☐	☐	✓	✓	✓	✓	✓	✓
PP Catalogs (DIASources, DIAObjects, DIAForcedSources)	☐	☐	✓	✓	✓	✓	✓	✓
PP Alerts (Canned)	☐	☐	✓	✓	✓	✓	✓	✓
PP Alerts (Live, Brokered)	☐	☐	☐	✓	✓	✓	✓	✓
PP SSP Catalogs	☐	☐	✓	✓	✓	✓	✓	✓
DRP SSP Catalogs	☐	☐	☐	☐	✓	✓	✓	✓

# Rubin data previews (cont.)

- **Data preview 0.1 (DP0.1)**

- image processing of simulated DESC DC2 sky survey, 5 years, 6 optical bands, wide-fast-deep (WFD) area of 300 deg<sup>2</sup> ([arXiv:2010.05926v3](https://arxiv.org/abs/2010.05926v3))
- processing performed by IN2P3 at CC-IN2P3 over 2020 & 2021 using the [LSST science pipelines](#) v19 (Dec '19)
- image and catalog products available via the [Rubin Science Platform](#) since June '21
- see documentation for DP0.1: <https://dp0-1.lsst.io>

- **Data preview 0.2 (DP0.2)**

- ongoing processing of same set of simulated raw images, with a more recent release of the pipelines v23 (Dec '21)
- performed independently both at Rubin Interim Data Facility (hosted in Google cloud) and at French Data Facility (FrDF), a.k.a. CC-IN2P3
- the products of this processing campaign are scheduled to be available for users in June '22



# Data preview 0.2 (DP0.2)

- ~20k visits, 1 exposures per visit, 189 detectors per visit
- Processing at Interim Data Facility (US) almost complete, ongoing at French Data Facility (FR)
  - several months of processing
  - required storage: 1.8 PB, including intermediate products that can be removed/archived when processing is complete

# DP0.2 preparation and tests at FrDF

- Documentation of the procedure for reproducibly creating a *butler* repository from scratch
  - <https://rtn-029.lsst.io>
- Getting familiar with the LSST "Batch Production Service" (BPS)
  - Framework for distributed pipeline execution
  - Understands what are the inputs and output of each task, generates a workflow (a directed graph of inter-dependent processing tasks) and submits it to an external Workflow Management System
  - A plugin is needed to interface BPS to the Workflow Management System which drives the execution of those tasks
- Explore existing Workflow Management Systems that could be plugged into BPS to run the processing workflows on our computing farm
  - Identification of [Parsl](#) as a good candidate (existing [plugin](#) developed by J. Chiang)
  - First tests of Parsl (with support from B. Clifford)

# DP0.2 preparation and tests at FrDF (cont)



We have been running hundreds of single-exposure and single-tract processing to test the various infrastructure components in order to:

- Validate the butler repository contents
- Validate LSST stack release updates and the software environment
- Test the workflow execution system and job management system
- Understand Parsl configuration and its scalability
- Probe infrastructure scalability (database, storage, jobs, etc.)
- Measure pipeline task resources usage and tune the batch jobs requirements

# DP0.2 preparation and tests at at FrDF (cont)



Test campaigns also allowed to validate our setup with various configurations:

- Migration of shared filesystem from GPFS to CephFS
- Migration of workload management system from GridEngine to Slurm
- Preliminary tests using dCache storage system
  - software component to allow the LSST Science Pipelines to use dCache as a butler data store developed by CC-IN2P3
  - several iterations of improvements and bug fixes
- Packaging and distribution of LSST software as Singularity containers
  - improve reproducibility
  - distributed under `/cvfms/sw.lsst.eu`
  - deployed at several sites, including SLAC (Rubin US Data Facility)

# DP0.2 production at FrDF

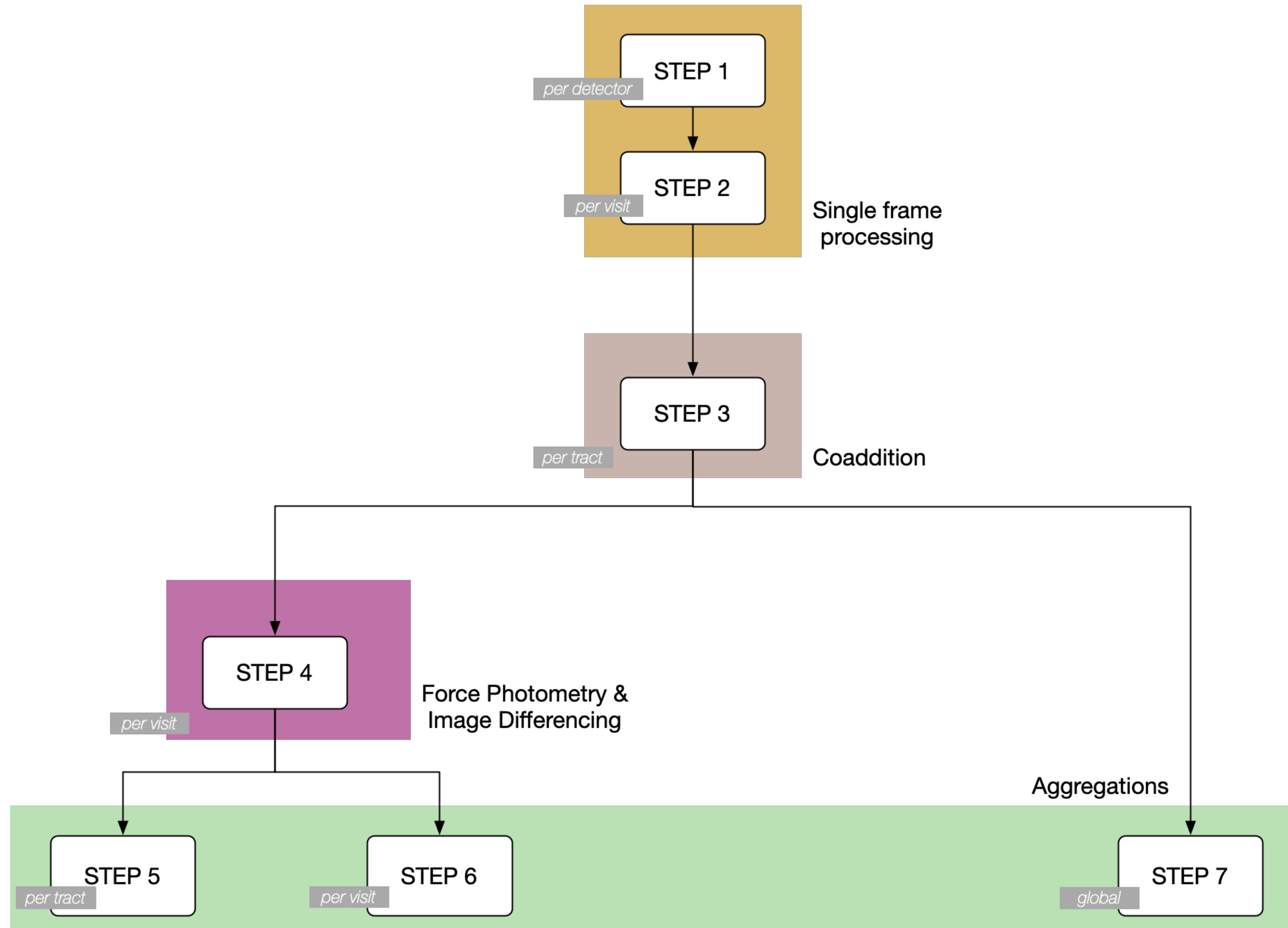
Production started mid-April:

- step1: completed with a few errors due to missing reference catalogs (detectors at the edge of the survey)
- step2: completed without errors
- step3: processing ongoing

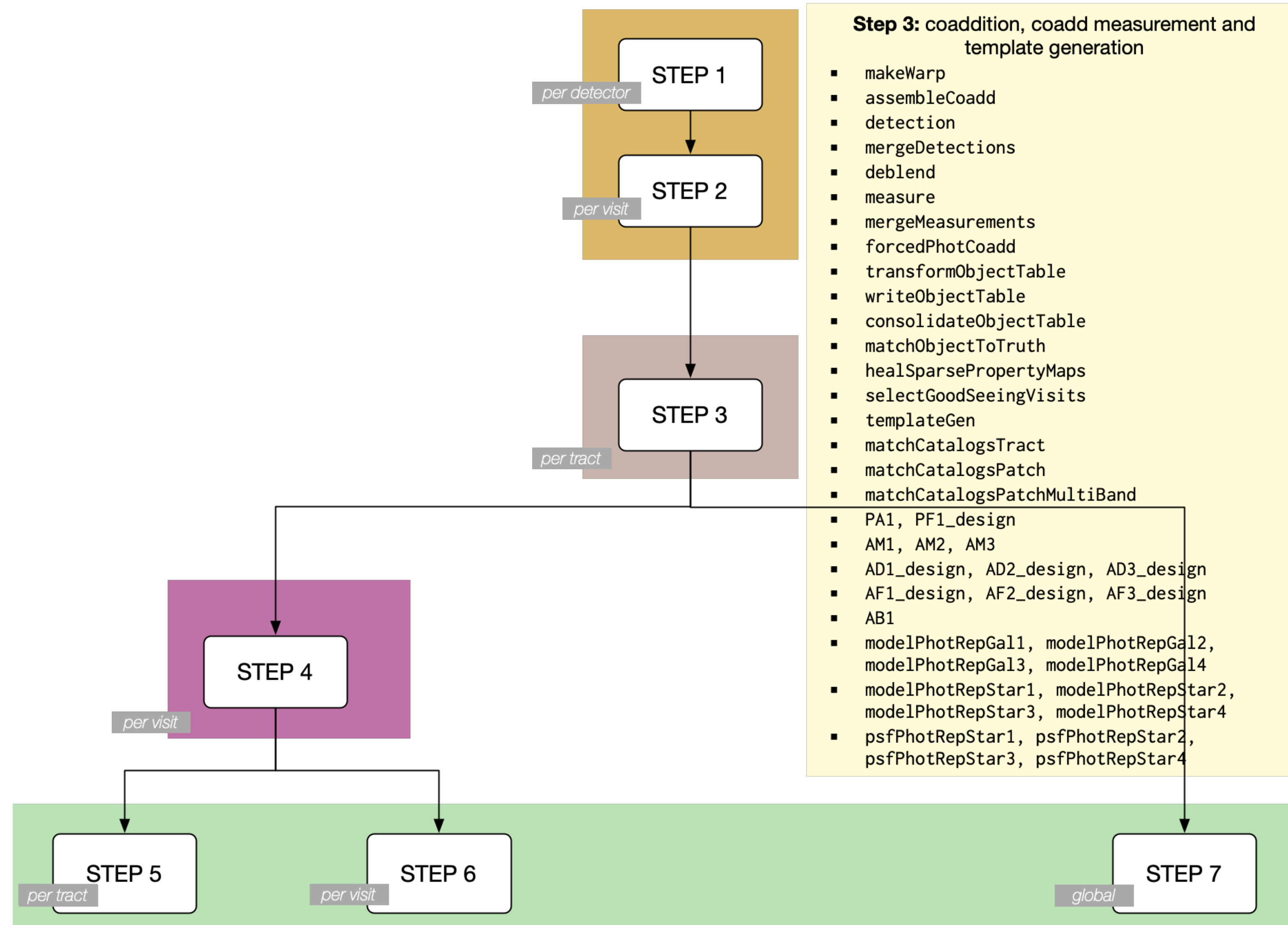
What comes next:

- Complete step3 which is probably the most challenging step (coaddition)
- Validate outputs
- Reprocess step1 to step3 using dCache as butler data store
- Complete step4 to step7 on both storage systems

## Rubin — Pictorial View of Data Release Processing



## Rubin — Pictorial View of Data Release Processing



# DP0.2: Lessons learnt

- Operation is more complex than just submitting jobs. Many issues are involved:
  - Splitting the full set of data to process them in smaller runs
  - Gathering tasks together in a consistent and scalable way
  - Parallelizing optimally the runs and jobs
  - Allocating optimal resources to the task
  - Detecting and characterizing failures
  - Reprocessing failed tasks when needed
- Additional tools are needed to have a good understanding of the processing:
  - Log management
  - Profiling of tasks
- Validation is also not trivial:
  - Tools being developed
  - What level of consistency with USDF should we reach?
  - How do we decide that we can process the next step?
- No big infrastructure issues up to now, but the memory requirement of some tasks and I/O activity could be concerning



- Our intention is to populate the catalog database (Qserv) with the products of DP0.2
  - either those produced at IDF or at FrDF
- Do we also need to have the image products at FrDF?
  - those produced at FrDF very likely not identical bit-by-bit to the ones produced at IDF
  - we should be able to import and ingest image products from IDF
- Programmatic access to butler repository requires access to the registry database
  - we need to identify mechanism for allowing multi-user access without risks
- Validation of products of data release processing is far from trivial

- Since a few years, we have been preparing IN2P3 contribution to data release processing
  - we are making progress but we expect this to be a never-ending learning experience
- We need involvement from more LSST-France scientists
  - it is time to get involved and collectively be prepared for exploiting the data
  - it takes some some time to get familiar with the LSST ecosystem, the tools and the data products
  - there is a lot of pedagogical material that has been recently developed by the project and the community