D2.5.3 Collection of final reports for mini-projects w/ DEV activities

WP2.5: Science Support

Project Acronym	LUSC-B
Project Title	UK Involvement in the Large Synoptic Survey Telescope
Document Number	LUSC-B-38

Submission date	8/JUN/2023
Version	1.1
Status	Published
Author(s) inc. institutional affiliation	George Beckett, Edinburgh; Mike Read, Edinburgh
Reviewer(s)	Ken Smith (QUB), Chris Frohmaier (Southampton)

Dissemination level	
Public	*This report may be distributed to any potentially interested parties.*

Version History

Version	Date	Comments, Changes, Status	Authors, contributors, reviewers
0.1	17/MAR/23	Initial version	MGB
0.2	22/MAR/23	Corrections and improvements	MAR, MGB
0.3	27/MAR/23	Corrections and removal of comments	MGB
1.0	8/JUN/23	Updated in light of reviewers' comments	MGB, MAR
1.1	30/JUN/23	Approved by LSST:UK Executive Committee	TMS

Table of Contents

Index of Figures

Index of Tables

1 Introduction

A unique element of the LSST:UK programme, and a particular advantage for UK astronomy, is the Development (DEV) work package, which enhances and extends the astronomy potential of the baseline Rubin Observatory software, services, and data products, for priority astronomy topics within the UK.

Some activities in the DEV work package need to interface with the LSST:UK Data Access Centre (DAC), so their outputs can be made available to science users during operations. A key activity in WP2.5 has been to engage early with the relevant DEV teams to develop, test and document how the DEV products will interact with the UK DAC and be supported by the DAC team during Operations. The output of these *mini projects* is a set of DAC-DEV interface-definition reports which define the interfaces between the DAC and each set of DEV products in sufficient detail to ensure the products integrate seamlessly into an astronomer's workflow, and to document where the responsibility for different aspects of the interface will lie (that is, with the DAC team or the DEV team). This work helps promote the impact of LSST:UK activities and contributes to the fulfilment of the science ambitions of the UK astronomy community.

The DAC-DEV interactions are built around four key objectives:

1. Familiarise the relevant DEV teams with the capabilities and functionalities of the UK DAC platform.
2. Identify key DAC interfaces with which DEV teams need to work, to integrate their outputs into the UK DAC.
3. Secure DAC resources and capabilities to host the DEV teams' outputs.
4. Test DAC-DEV integration, based on precursor data sets or performing other representative tests, to validate the selected approach and interfaces.

From the Phase B DEV work package, two activities have been identified whose products (software and/ or datasets) are to be integrated into the LSST:UK DAC:

- WP3.5: LSST and Near-IR Data Fusion
- WP3.11: Crossmatching and Astrometry at LSST Depths

At the time of writing, these DAC-DEV mini-projects are close to complete, and the outcome of the work is documented in two technical reports introduced in Section 2.

Sustainability has been an important consideration for both DEV activities, as the (DEV) teams who are developing the products are unlikely to continue (in their current roles) into Rubin Operations. Some of the responsibility for ensuring sustainability will rest with Recipient Groups (through the In-kind Agreement) and this is briefly discussed in the in-kind manual [1]. However, some of the responsibility will lie with LSST:UK (in particular, the DAC team) and, to help us address this, the DAC team has engaged with the Software Sustainability Institute (www.software.ac.uk). The Institute helped us to develop an LSST:UK software-development plan and have helped us to deliver an intermediate-level software-engineering training to the teams. We have maintained a dialogue regarding DAC-DEV Interface Requirements documents for advice and guidance.

1.1 Glossary of Acronyms

CSD3	Cambridge System for Data Driven Discovery
DAC	Data Access Centre

DEV	[LSST:UK] Development project
ESO	European Southern Observatory
HPC	High-performance Computing
RSP	Rubin Science Platform
SSI	Software Sustainability Institute

2 Highlights and Exceptions

2.1 WP3.5 LSST and Near-IR Data Fusion

The Work Package 3.5 team is developing an extension to the LSST Pipeline software (commonly referred to as the LSST Stack), to support joint processing of LSST pixels together with pixel data from the ESO VISTA Public Surveys, to produce merged optical and near-infrared catalogues. Joint processing of LSST and VISTA data will significantly extend the capabilities of the baseline LSST Data Releases and has already been accepted as part of the LSST:UK in-kind contribution to the Rubin Observatory, with the Galaxies Science Collaboration and the Active Galactic Nuclei Science Collaboration as primary and secondary Recipient Groups, respectively [1].

The WP3.5 in-kind contribution focuses on high-latitude, extragalactic surveys (from VISTA). Notably VHS, VIKING, and VIDEO (plus, possibly VEILS), which all overlap with the LSST survey. These VISTA surveys are all due to be completed in advance of the beginning of Rubin Observatory operations.

For each LSST Data Release, WP3.5 will provide a collection of VISTA deep coadd images, which coincide with LSST images and which extend the range of available observing *bands*. These will be supplemented by object catalogues in the LSST schema and new measurement catalogues that exploit information from the combined VISTA-LSST coadd image sets. These products will be made available to the Rubin Community via the UK DAC.

Details of how the extended pipeline is run and how outputs from the pipeline are ingested into the UK DAC is provided in an LSST:UK technical report [2] and summarised here.

A typical workflow with the LSST Stack involves four stages:

1. Ingest images ready to be processed, manipulated, or analysed with the LSST Stack.
2. Define a workflow, using Stack-specific tooling.
3. Run the workflow on a suitable resource.
4. Interrogate the results by either working directly with the data or publish the results into a Rubin Science Platform for wider consumption.

A typical WP3.5 run of the pipeline is illustrated in Figure 1. The workflow has been tested at scale using the Hyper Suprime-Cam (HSC) survey as a proxy for LSST. Based on this, a full run fusing an LSST data release with the VISTA wide surveys (VIKING, VHS and VIDEO) will take around 3.2 million CPU hours and involve around 2 Petabytes of working data. This pipeline would be run for each of the early data releases, with each run expected to require a similar amount of resource (unless the LSST Pipeline changes significantly). The workflow is large and complex, so specialised software – namely, the Parsl workflow manager and the LSST Batch Processing System – are used to automate the execution of the workflow.

The large resource requirements of the pipeline mean a processing campaign can only reasonably be completed on a national-scale high-performance-computing system. The DiRAC HPC service, CSD3, available via IRIS, has been demonstrated to be a suitable platform on which to process the data. An LSST-scale run is expected to take up to four weeks to complete on a system such as CSD3.

At the time of writing, a wide run of VISTA (based on VIKING, VHS, and VIDEO surveys) fused with HSC data is being finalised for ingestion into the UK DAC, from where it will be made available to a small number of interested scientists for validation.

The team will continue to develop the pipeline, for fusing VISTA data with LSST, in Phase C. Given that the LSST Stack has yet to be finalised, it is likely further refinements of the pipeline will be made during that phase.

2.2 WP3.11 Crossmatching and Astrometry at LSST Depths

The WP3.11 team's in-kind contribution to the Rubin Observatory involves making crossmatches between the LSST and appropriate ancillary surveys (see Table 1 for a topical list) This is done using a bespoke algorithm, developed by the WP3.11 team, which has better than usual accuracy in crowded fields. The WP3.11 team are also making the software implementation of the crossmatch algorithm, which is called *Macauff* [3], available to the astronomy community.

Ancillary Surveys	Catalogue Size (Object Count)	Format
Gaia	3 TB (1.7 Bn)	SQL
VISTA (multiple surveys)	10 TB (5 Bn)	SQL
WISE	1–3 TB (1.6 Bn)	SQL
VPHAS	1 TB (1 Bn)	SQL
Spitzer (surveys tbc., though likely Glimpse)	250 GB (300 Mn)	SQL/CSV
SkyMapper (*)	1 TB (0.3 Bn)	SQL
Schmidt surveys (*)	<1 TB (1 Bn)	SQL
2MASS (*)	<1 TB (0.5 Bn)	SQL

Table 1: List of ancillaries for which crossmatch with LSST DRs is valuable (* indicates a potential additional survey, which will only be processed if time and resources permit). For comparison, each LSST DR is expected to catalogue 35 billion objects.

The in-kind contribution has been endorsed by the Transient and Variable Stars (TVS) Rubin Science Collaboration and the Stars, Milky Way, and Local Volume (SMWLV) Rubin Science Collaboration. The WP3.11 team engage with the Rubin Crowded Field Joint Taskforce, which is a TVS-SMWLV joint taskforce.

Details of how crossmatch products are generated and integrated into the DAC are provided in an LSST:UK technical report [3] and summarised here. At the top level, crossmatching between LSST objects and an ancillary survey involves a three-phase workflow:

1. The relevant LSST object catalogue and the third-party catalogue are pre-processed into a canonical format, called *skinny tables*, suitable for ingestion into Macauff.
2. A Macauff workflow is submitted to a suitable IRIS resource, in which the survey inputs are analysed, creating several new dataset that describes identified correspondences between the two surveys – using two different approaches (astrometry-based and photometry-based) plus identifying objects for which no crossmatch exists.
3. Once validated, the crossmatch dataset is ingested into the UK DAC and registered with the Rubin Science Platform (RSP) via an IVOA TAP (Table Access Protocol) service.

A visual representation of the workflow is provided in Figure 2.

The workflow has been tested at scale using representative survey data from WISE, as a precursor for LSST, and Gaia, as an interesting ancillary survey with good sky overlap with WISE (and LSST).

At LSST scale, Macauff requires significant computing resources to complete in a timely manner and the DiRAC HPC service at Cambridge (CSD3) has been identified as an appropriate service. To exploit the capabilities of CSD3, a distributed-memory high-performance computer, the Macauff implementation needed to be parallelised and optimised. At the conclusion of this work, it is estimated a full crossmatch of an LSST Data Release object catalogue against one ancillary survey of interest could be completed in around one week of CSD3 time, using around 240,000 CPU hours and 100 Terabytes of working storage. This means that crossmatch catalogues for the selected ancillary surveys should be produced and published in the UK DAC (ideally) within a month of an LSST Data Release being published.

At the time of writing, the WISE-Gaia crossmatch catalogues are being ingested into the UK DAC from where they will be used by a PhD student at the University of Exeter to support their research. Once complete, the experience of the student will be captured to inform future enhancements to the workflow and the presentation of the data within the DAC.

The WP3.11 team's work will be extended in Phase C to create:

- super-match catalogues, which identify correspondence between LSST and multiple ancillary surveys, in a single catalogue.
- a catalogue of rare (thus, potentially, very interesting) objects.

3 Conclusions

At the time of writing, there is a good understanding of how both DEV activities will provide User-generated Products for the UK DAC. Having identified key interfaces in their respective interface-definition documents ([2] and [3]) provides a way to identify issues that may be caused by any changes to the supporting (Rubin) software or due to changes in the science requirements of the recipient groups. We also have scalable test cases for both workflows, which can be exercised periodically to validate the infrastructure as it is configured in the run-up to telescope operations, in late 2024.

The resource requirements of both DEV activities are considerable and require support from the IRIS programme. The estimated resource requirements have been factored into the forward estimates that we provide to IRIS and are confirming to be within the anticipated capacity of the IRIS programme.

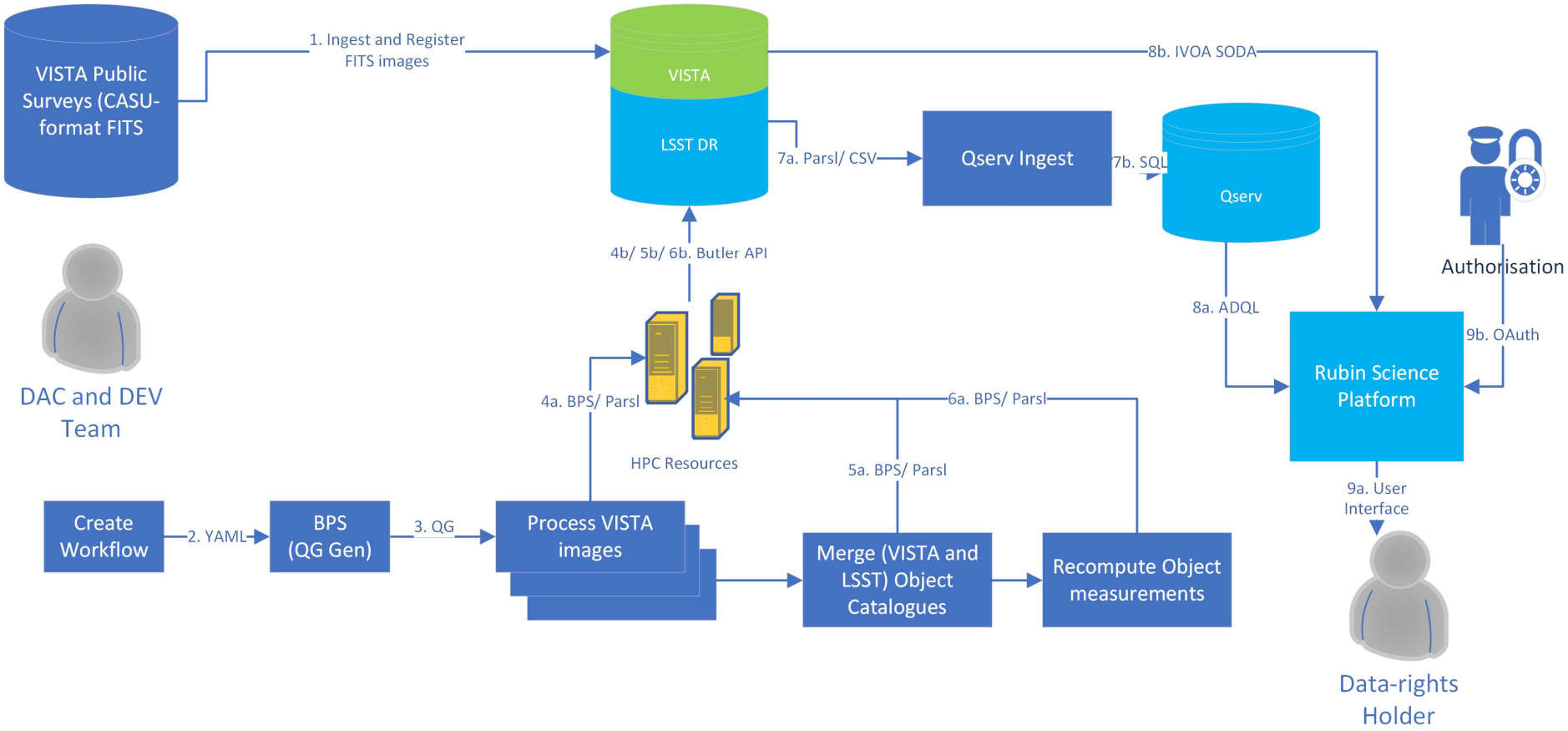

Figure 1: High-level Overview of Workflow to Generated Fused Data Products

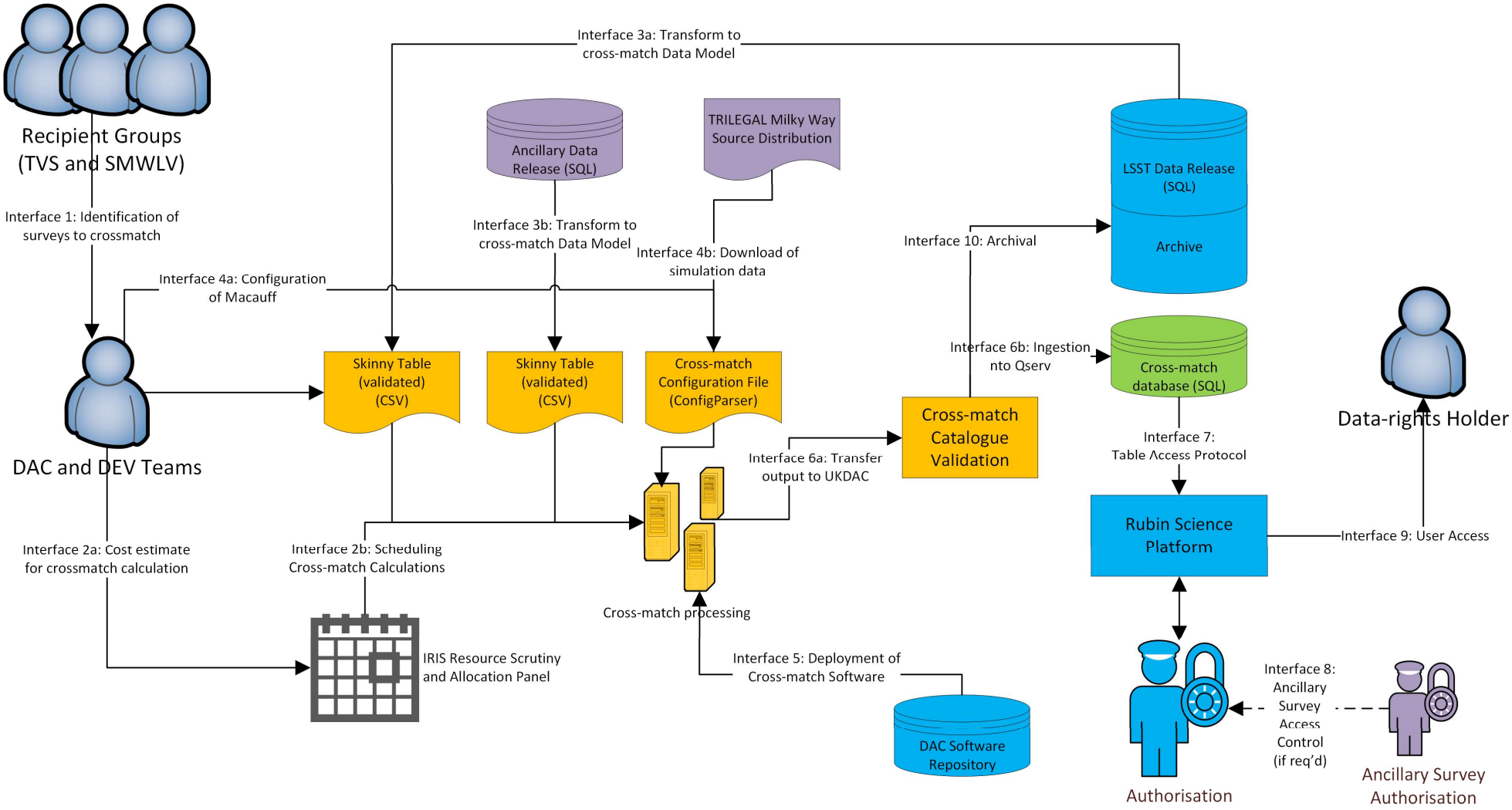

Figure 2: Workflow for generating a Crossmatch Catalogue.

4 References

[1] Phil Marshall, et al., *RDO-41: Manual for In-kind Contributors and Recipients*. Rubin Observatory https://docushare.lsst.org/docushare/dsweb/Get/RDO-41 (July 2021).

[2] George Beckett, et al., *WP3.5 DAC-DEV Interface Requirements*, LSST:UK Phase B Technical Report, LUSC-B-22 (March 2023) https://lsst-uk.atlassian.net/wiki/download/attachments/766869533/WP3-5_DAC-DEV_interface_definition.pdf

[3] George Beckett, et al. *WP3.11 DAC-DEV Interface Requirements*, LSST:UK Phase B Technical Report, LUSC-B-20 (March 2023) https://lsst-uk.atlassian.net/wiki/download/attachments/766574618/WP3-11_DAC-DEV_interface_definition.pdf