



Eugene Magnier <eugene.magnier@gmail.com>

ApJS - Referee's report on AAS17743

steven.crawford@aes.org <steven.crawford@aes.org>
Reply-To: steven.crawford@aes.org, peer.review@aes.org
To: eugene@ifa.hawaii.edu
Cc: steven.crawford@aes.org

Wed, Jun 12, 2019 at 1:42 PM

June 12, 2019

Dr. Eugene A. Magnier
University of Hawaii at Manoa
Institute for Astronomy
[2680 Woodlawn Drive](#)
[Honolulu, HI 96822-1897](#)

Title: The Pan-STARRS Data Processing System

Dear Dr. Magnier,

We have received the referee's report on your above submission to The Astrophysical Journal Supplement Series, and it is appended below. As you will see, the referee thinks highly of your work and has only a few suggestions for changes.

When you resubmit the manuscript, please include a cover letter in which you outline in detail the specific changes you have made in response to each of the referee's comments.

Click the link below to upload your revised manuscript;

<https://aes.msubmit.net/cgi-bin/main.plex?el=A6KO1CZq3A7Hlp4J3A9ftdHbVhx5lycbK5x0IHbt0zgZ>

Alternatively, you may also log into your account at the EJ Press web site, <http://apj.msubmit.net>. Please use your user's login name: emagnier. You can then ask for a new password via the Unknown/Forgotten Password link if you have forgotten your password.

Reviewers find it helpful if the changes in the text of the manuscript are easily distinguishable from the rest of the text. Therefore we ask you to print changes in bold face. The highlighting can be removed easily after the review.

The AAS Journals have adopted a new policy that manuscript files become inactive, and are considered to have been withdrawn, six months after the most recent referee's report goes to the authors, provided a revised version has not been received by that time.

If you have any questions, feel free to contact me.

Sincerely,

Dr. Steven Crawford
AAS Scientific Editor

Referee Report

Reviewer's Comments:

General Notes

This is a well-written and important technical paper that succeeds admirably at what I consider the most important goal of any pipeline paper: providing a description of the processing steps that are relevant for downstream science users (in this case, by providing the big picture that ties together a number of more detailed papers). With only a handful of minor cleanups (see detailed notes below), I think the paper is ready for publication, and most of my comments represent ideas for improvement that I hope the authors will consider (but should not feel obliged to act on).

My only general concern is that the paper often misses the opportunity to pass on lessons learned to the developers of

future pipelines, and this makes much of the detailed description of how the PS1 systems work (particularly in Section 5) feel like it belongs more in operator documentation rather than an article like this one. I suspect a small amount of additional historical context - how different systems evolved over the course of the survey - and commentary on what worked well and what was a regular pain point would go a long way.

In particular, the described system seems to involve a both fair amount of duplication (e.g. multiple databases, sky-tiling systems, and task orchestration layers) and a number of in-house solutions to what seem like fairly general problems (the DVO database and especially the pantask/opihi system stand out in this regard). This is not intended as criticism; I am quite aware that there are many good reasons for both duplication and keeping central components in-house, from deliberately keeping components loosely coupled to taking into account the often-brief shelf-life of off-the-shelf solutions, especially as compared to the duration of a major astronomical survey. But describing *which* of many potential reasons actually played a role in each of various design choices (and which, if any, look less good in hindsight) would make the paper much more interesting.

Detailed Notes

Section 2.4

Is the Distribution and Publication system mentioned in the text supposed to be part of Figure 1, either as an umbrella term or a (missing?) component?

Section 3.1

This is by no means necessary, but I'm curious to see a table or discussion of what fraction of jobs of various types failed with bad "quality". In other words, how much data could you not get through the pipelines at all, and what was the most sensitive step?

Section 3.3

Running "Registration" only once for each exposure would seem to prohibit re-running "burntool" after updating the algorithm for that - and I'm guessing you didn't get that fully stabilized until after you had already processed some images and learned from the experience. How did that work?

Section 3.5

Why a 3rd-order polynomial from chip to focal plane? Wouldn't an affine transform have been sufficient (and more than that degenerate with the focal plane to sky transform)?

What makes the masks generated in this step "dynamic"? Are they generated wholly from the reference catalog (i.e. predicting where a ghost will appear based on the position of a bright star)? It seems like the CAMERA step does not utilize any of the pixel data (just the pixel-level masks from CHIP). Is that correct?

Section 3.8

Is the selection of which warped images go into a stack driven by human operators, or are there automated systems to launch these jobs, too?

Section 3.10

How much of the PSF-convolved galaxy models do you re-fit in forced photometry? If you're fitting more than just the amplitude at that stage, and considering each exposure as independent, you're potentially throwing away a lot of S/N (at least in the many-exposure limit), even if you average later. If you're just fitting the amplitude, the structural parameters are still going to be the ones affected by poor PSFs in the stack.

Section 3.11

transient source -> transient sources

Section 4.1.3

I was confused when first encountering the word "files" here because up to this point I had been thinking of the DVO as just another MySQL (or other SQL-ish) database, and I wasn't sure what kind of files were being referred to. I think it'd be helpful to briefly describe the overall architecture of the DVO as (mostly?) spatially sharded files at the beginning of section 4, even if the details of the partitioning aren't described until 4.1.3.

Missing punctuation in parenthetical HST GSC reference?

Section 4.1.4

There's some inconsistency here between "detID" and "det_id" (same for "image"), both referring to measurement IDs in DVO. If those are supposed to be meaningfully different, I'm confused.

Section 4.2

I tend to associate the term "übercal" specifically with the SDSS version of the algorithm that coined the term, and think it probably should be referenced here even if the actual algorithms used are only vaguely similar.

Section 4.3

Is the PSPS database another spatially-shared, file-based database using custom technology, a MySQL database like the Processing Database, or something else? I assume the same system is used at both IFA and MAST?

Section 5.1.1

Apparent typo or missing text: macro ex- its job successfully".

Section 5.1.4

"responsible to" -> "responsible for"

Section 5.2

> Pairing warps together is simplified by the observing strategy in which the same pointing is observed multiple times in a night. By limiting to warp-warp pairs from the same pointing, the problem is significantly reduced from the arbitrary case.

This (as well as the following paragraph) seems to imply that you typically generate differences between images taken in the same night, which of course limits you to detecting only very short-timescale transients and fast-moving objects. I suspect that's just not what you intended to imply, or is the nightly processing really not supposed to find e.g. supernovae?

Section 5.2

Are the `projection\_cells` described here the same as or related to the DVO partition cells of 4.1.3, or the RINGS.V3 skycells of 3.7?

This is a more general concept, but it came to a head in this section: I found the use of so many notation styles for different concepts more distracting than helpful. I think I was able to infer that small caps were used for processing stages and non-bold italics were used for database tables, but it wasn't clear why some other stages were written in fixed-width mixed case instead (were these scripts, rather than stages?), or what the use of bold-italic meant (everything else?). I'd recommend either adding a notation legend paragraph early in the paper or just cutting down on the number of styles used.

Section 5.3

Was Nebulous just used by the orchestration levels like pantasks, or was it used within the Perl scripts and C programs that constitute the algorithmic steps as well?

Was the database used by Nebulous integrated with the Processing Database at all (or even part of the same server)?

It's a bit strange to first encounter what seems like a core part of the data access system this late in the description, given that it would have needed to be updated by all of the processing steps mentioned early. This would of course make more sense if Nebulous is in fact used by the lowest levels of the pipeline and hence a Nebulous database entry is created whenever a file is written to disk.