# Nancy Grace Roman Space Telescope (Roman) Technical Report

| Title: Roman WFI Co-add catalog schema | Doc #: Roman-STScI-000766, SE-01<br>Date: 7 May 2025<br>Rev: - |
|---|---|
| Authors: Henry C. Ferguson, Eddie Schlafly, Tyler Desjardins, Dario Fadda, Adrea Bellini, Nadia Dencheva, Larry Bradley, Brian McLean<br>Phone: 410-338-5098 | Release Date: 11 June 2025 |

## 1. Abstract

The Science Operations Center (SOC) for Roman at the Space Telescope Science Institute (STScI) will be adding together, and re-projecting onto a rectified sky grid, the individual exposures from the Roman Wide Field Instrument (WFI). These co-added images are called "level 3 (L3) data products," and consist of several variants constructed at different times (prompt vs. data-release) and different depth (visits, subsets, and full) **(Ferguson & et al., 2025)**. The SOC pipeline will then construct catalogs of astronomical objects in these L3 images. The catalogs and associated ancillary data constitute the L4 data products. (There are also L4 catalog data products that will be generated from L2 images; these are not discussed in this document.)

This technical report summarizes the decisions for what fields (i.e. measurements) to put into the first implementation of the L4 catalogs to be made from the L3 images, both for prompt and data-release products. To preserve information for future changes, the accompanying excel spreadsheet also lists the fields that have been considered, but which are NOT planned for inclusion into the first implementation of the catalogs. These can be considered for future releases of the SOC pipeline. This report does not include details of how the measurements are made; that discussion is in a separate technical report that outlines the co-add catalog strategy.

## 2. Introduction

For *Hubble* and the *James Webb Space Telescope*, most of the measurements to produce catalogs have been carried out by individual investigation teams. Many users do some custom processing and stacking of the images before making their catalogs. Most observations taken with *Hubble* and *Webb* are dithered – with the telescope pointing moved by a small amount between individual exposures. This mitigates the impact of any detector artifacts on the final stacked images that are created by coadding all the exposures – or specific subsets of the images – in a given photometric band. The wide variety of observing strategies for *Hubble* and *Webb* has made it impractical to develop a single pipeline that can produce image stacks that would be considered of "science quality" for many of the programs. Among the reasons this is impractical are: (1) the narrow fields of view make image registration challenging and while advances have been made to the process, precisely registering the images before stacking often requires human intervention; (2) the observation structure specified in the proposal planning system does not always map in a straightforward way to a specification of which exposures to combine into which kinds of image stacks; and (3) different observing strategies often require different approaches to outlier rejection and artifact removal.

Roman differs in that it is a survey mission with a wide field of view and a very high data rate. The wide field of view and stable focal plane means the SOC pipeline can automatically register the individual exposures sufficiently well to meet the science requirements using *Gaia* [MOU1]reference stars. With a smaller variety of observing modes and observing strategies, it is practical to develop an approach for making stacks of various depth that will probably meet the needs of most science. The SOC will of course need to be very responsive to community feedback to ensure that this is the case throughout the mission.

Starting with science-quality stacks, it is now feasible for the SOC pipeline to produce science-quality catalogs. These are intended to have astrometric, photometric, and morphological measurements close to the optimal precisions possible given the underlying data.

However, they are not specifically aimed at meeting the science requirements of Roman Core Community Surveys. The Project Infrastructure Teams (PITs) have been created to achieve that. These teams will be doing custom processing on the Roman data that is focused on their specific science goals, as well as making use of data from other facilities (e.g. Rubin) to augment the Roman data.

The general community is likely to use a combination of the PIT-provided data products and the SOC-provided data products. The SOC is formally responsible for producing general community-oriented products from the WFI imaging. The PITs are responsible for ensuring that Roman achieves the core science objectives specified in the requirements, but are not responsible for providing general-purpose data products.

With billions of stars and galaxies to measure and a wide variety of quantities that could be measured, it is important to try to optimize what to put in the SOC-generated catalogs. Smaller tables mean that catalog generation is faster, data storage is less expensive and database queries are faster. On the other hand, if a quantity of interest is not easily derived from columns already in the catalog, it is better for the SOC to make the measurement while the pixels are in memory rather than require teams that need the measurements to do this independently.

These trades have been discussed internally at STScI, in multiple meetings of the Software Working Group, at a SOC/SSC Technical Interchange Meeting, and at several Roman Science Quarterlies. Based on that input and feedback, this document puts forward a definitive set of fields for the SOC baseline WFI catalogs.

The L3 images are relatively small "skycells" that overlap adjacent sky cells by a small amount. Each skycell is roughly 5000x5000 pixels[MOU2][HF3] when dithers allow oversampling to 0.05 arcseconds per pixel. **(Fadda & et al., 2024)** (The algorithm for deciding the pixel scale has not been finalized; the 0.05 arcsecond pixel scale is probably appropriate when there are at least 3 dithered exposures, for most of the available dither patterns.) The initial generation of the catalogs is entirely atomic within each skycell: it does not require any of the adjacent skycells or even any metadata about them. The output of this cataloging step from romancal is an individual-skycell catalog in parquet format. Until recently, the format for these files was to be the Advanced Scientific Data Format (ASDF). However parquet is now being adopted for wider compatibility with general data-science tools. To be clear that these files – when they are generated by romancal – are for an individual skycell, we will refer to these as **skycell-parquet** files in the remainder of this document.

The catalogs for the individual skycells are merged downstream, removing duplications between the overlapping regions. The current algorithm for identifying the duplicates (to be removed) is as follows. For each skycell, we have designated a number of overlap pixels with the adjacent skycells (skycell_border_pixels = 100) and the size of each skycell (5000 pixels = 4.16 arcminutes when using 0.05" pixels). These skycells are on a rectangular grid, all projected onto a common tangent plane over a projection region (roughly two degrees on a side). Each projection region is defined by a range of RA and Dec. A detection is primary if its position is within the RA / Dec boundaries of the projection region it is associated with, as well within the core region of the skycell (i.e. more than skycell_border_pixels away from the edge of the skycell). Otherwise the object is flagged as being in the overlap region. In the core region, the current skycell has the best measurement, while if the object is centered in the overlap region, the neighboring cell will probably a measurement that is at least as good, or better. In the schema we have two ways to encode this flag: we include it as the highest-order bit in the 64-bit flagged_spatial_index, and we include it as the lowest-order bit in warning_flags. The reason for doing both is that having this bit set in the flagged_spatial_index (which can act as a sort of an object ID) is a clear indication that this object probably has a better measurement in a different skycell.

The prompt versions of the L3 images are produced as the data arrive. **(Ferguson & et al., 2025)** For the prompt L3 images, the current plan is for these to be co-additions of just the images within a single visit. For data releases, the co-addition will depend on the survey observing strategy, but in general will include multiple prior visits. The L4 prompt catalogs will be generated for each L3 image as soon as the L3 image is available. These images will be completely independent for each photometric band. There may be different numbers of sources in each band because the detections are entirely independent.

The data-release L3 images will have been reprocessed with consistent calibration across the

entire survey dataset. The L4 data-release catalogs will include multi-band point-spread-function-matched (PSF-matched) photometry based on a single detection image. Details of the detection strategy are not specified in this document because they are still being discussed.

The core of this document is a spreadsheet that specifies which measurements should be included in the baseline versions of these catalogs. The spreadsheet lists specific names, data types and units, and attempts a consistent naming convention for the different quantities.

## 3. Data Products

The catalogs are first created as parquet files – and there may be several of these for each WFI image, depending on the software implementation. Some of the fields that are useful to astronomers are a bit redundant – in that they are just different representations of the same information. These redundant fields need not be in the initial set of skycell-parquet files because they could be generated downstream – either when the data are ingested into The Mikulski Archive for Space Telescopes (MAST) databases or even on the fly when the database is queried. This document identifies the derived quantities and makes a recommendation of which should appear in the skycell-parquet files, which should appear in the MAST databases and which should appear in both.

The individual-skycell parquet files will be used by the Science Support Center (SSC) as input to the spectroscopy data processing. However, for general users, the MAST database versions of the catalogs will be more useful, because they will have been merged across all of the skycells and projection regions that make up each survey.

From the perspective of software maintenance, there is an important advantage to isolating these downstream "derived quantity" calculations in a separate module from the one that requires access to the pixels: that software takes tables as input and creates tables as output and therefore requires only tables for testing and operations. On the other hand, for some quanitities the raw version may be so rarely used that it makes sense to do the conversion to the derived quantity on the fly and save only that version when doing the initial computations. These are listed as Raw→Derived in the attached spreadsheet.

There are other "software architecture" issues that will influence the details of the skycell-parquet files. There may well be separate software modules for detection, for photometry, for different kinds of shape measurements and/or for photometric redshifts. (This facilitates software maintenance as well as parallelization of steps.) The L3 files for all the individual bands need not all be in memory at the same time. It is possible that different modules will write the results out to separate intermediate tables. These intermediate tables will be deleted when – at the end of processing the individual skycell – the tables are merged into a single skycell-parquet file, with the schema described in this document. For data releases, the skycell-parquet files will include the measurements from all of the available bands. Where measurements are made in all the bands, the field name includes a "_<band>" string (f062, f087, f106, f129, f158, f184, f213 or f146). For prompt processing, there will be separate skycell-parquet files for each photometric band. The names of the columns will therefore not include the "_<band>" field for the prompt

catalogs.

The data-release skycell-parquet files are to be merged across the full survey area into a MAST catalog database table. At this stage, duplicate entries in the overlapping regions of neighboring skycells can be identified. It is not practical to do this de-duplication prior to merging. Therefore the skycell-parquet file versions of the data-release catalogs will have separate records for the same physical object in these overlapping regions. In merging the files, an algorithm will determine which record is the "primary" record. The other records will be flagged and/or moved to a separate database table. In any case, this processing happens downstream of generating the skycell-parquet files.

While the algorithm for deciding which record is primary is deterministic, it cannot be run before all the catalogs of the overlapping skycells are complete. Because this processing is asynchronous, it is impractical to do the de-duplication before finishing the catalogs for all the skycells. The SSC will need to take this into account, as they are planning to drive their Grism two-dimensional data processing (G2DP) spectral extraction based on the skycell-parquet versions of the catalogs.

## 4. Fields to include or not include in the Baseline catalogs

The full list of fields considered for inclusion in the catalogs is attached as an excel spreadsheet. For convenience, a few of the spreadsheet columns for the subset of fields that are included in the baseline (for either prompt or data releases) are included in Table 2.

The full excel spreadsheet specifies the following, for each field:

- Priority: an assessment of the relative importance of the different fields (1 is highest).
- The next four columns indicate a decision of whether or not to include each field in the following ("Y" indicates inclusion):
    - Prompt: The prompt catalogs.
    - DR: The data-release catalogs.
    - Parquet: The skycell-parquet versions of the catalogs initially produced by romancal.
    - MAST: The merged versions of the catalogs in the MAST databases.
- Type: The data type.
- Unit: Units on this quantity.
- Category – one of the following:
    - Detection – measured during the detection process.
    - Photometry – a photometric measurement.
    - Position – a positional measurement.
    - Name – a name.
    - Index – an integer (or long integer) that may be used as a database index.
    - Shape – A morphological measurement.
    - Neighbors – A measurement using neighbors.
    - Photo-z – photometric redshift related information.
    - Sersic – Galaxy profile fitting.
    - Adaptive Moments – Galaxy adaptive moments.

    - Asinh – An alternative representation of fluxes.
- Raw/Derived – One of the following:
    - Raw: Needs access to the pixels to compute this.
    - Derived: Can be computed from another field in the catalog.
    - Raw→Derived: Needs access to the pixels, but do the conversion on the fly and save only the derived value, not both.
- Missing: What to use for missing data.
- Undefined: What to use for undefined data (e.g. magnitudes, where fluxes are less than zero).
- Likely use in query qualification: An assessment of how often this field might be used as part of a search query.
- Format: Suggested print format (using python style format specification).
- Min: Minimum valid value.
- Max: Maximum valid value.
- Short Description: A description of the field suitable for use as the description field in the data model.
- Description: A longer description of this field. This is to provide more detail for the algorithm development or documentation development. It is not intended for use beyond this document.
- Comments: Further comments about this field or algorithms.
- Algorithm: Comments about possible algorithms.
- Code: Pointers to existing code.

The spreadsheet can be filtered to show various subsets, based on the column headings. Please note that the unfiltered version of the spreadsheet includes many fields that have been suggested but are not intended to be included in the baseline version of the catalog.

For convenience, a few of the spreadsheet columns for the subset of fields that are included in the baseline (for either prompt or data releases) are included as a table below.

## 5. Open issues

There will almost certainly be some changes to the schema as the software for making the catalogs is further developed. Known open issues are listed below.

*Table 1: Open issues (mostly in how to compute various quantities in the catalogs).*

| Field | Issue |
|---|---|
| cxx,cxy,cyy | Treatment of pixels with negative fluxes in computing these not yet specified. |
| primary_key | Specification of how to compute this is not finalized. |
| association_key | This depends on a concept for a set of secondary tables. That concept might evolve. |
| healpix_id | Needs to be fully specified (Nside, nested vs. ring). |
| image_flags, | We can expect more flag bits to be set as we figure out what to flag. |

| | |
|---|---|
| warning_flags | |
| ra, dec | Algorithm for recommending the best coordinates might evolve. |
| e(b-v) | Source of reddening data still TBD. We need this for photo-z. |
| psf_gof | The goodness of fit metric might evolve. |
| *<band>m<band>* | Need to specify how to choose the reference band for arbitrary surveys. |
| exptime_proxy | The specific exposure-time proxy is still TBD. |
| nn_label, nn_distance | Currently computed within a skycell. This may cause confusion when the catalog is merged into a single database table in MAST. |
| is_extended | Algorithm for populating this field is not yet specified. |

*Table 2: Abbreviated schema for the fields to be included in the baseline catalog.*

## Acronyms

| Acronym | Definition |
|---|---|
| ASDF | Advanced Scientific Data Format |
| DMS | Data Management System |
| G2DP | Grism 2-Dimensional Processing |
| L1 | Level 1– Reformatted telemetry with metadata |
| L2 | Level 2 – Individual exposures with basic calibration applied |
| L3 | Level 3 – Reprojected onto a tangent plane and possibly co-added |
| L4 | Level 4 – Catalogs and associated ancillary products |
| L5 | Level 5 – Community contributed data products |
| MAST | Mikulski Archive for Space Telescopes |
| NASA | National Aeronautics and Space Administration |
| PIT | Project Infrastructure Team |
| PSF | Point-Spread Function |
| SOC | Science Operations Center |
| SSC | Science Support Center |
| STScI | Space Telescope Science Institute |
| WFI | Wide Field Instrument |

## References

Fadda, D., & et al. (2024). *The Roman Tesselation of the Sky Sphere.* Roman-STScI-000708.
Ferguson, H. C., & et al. (2025). *Roman Reprocessing and Data Release Strategy.* Roman-STScI-XXXXXX.