Nancy Grace Roman Space Telescope (Roman)
Technical Report

Title: Roman Science Operations Center Wide Field Instrument data processing summary	Doc #: Roman-STScI-00NNNN, SC-01 Date: Rev:
Authors: Henry Ferguson, Megan Sosey, Greg Snyder, Erica Kolatch Phone: (410) 338-5098	Release Date: LEAVE BLANK Draft

Abstract

The Roman Space Telescope's Science Operations Center (SOC) at the Space Telescope Science Institute (STScI) is responsible for many of the data processing steps for the Wide Field Instrument (WFI), providing a data-analysis platform for astronomers to inspect the data and carry out basic data analysis, and providing archived Roman data through the Mikulski Archive for Space Telescopes (MAST). This document provides a high-level description of the Data Management System (DMS) WFI data processing and data products as well as an overview of the Roman Science Platform and MAST services. This is intended to provide a snapshot of the plans as they stand at the beginning of calendar year 2022.

Operated by the Association of Universities for Research in Astronomy, Inc., for the National Aeronautics and Space Administration under Contract #80GSFC19C0054

Table of Contents

1 Introduction 3
2 Pipeline Data Products 4
2.1 Survey Data Releases 6
2.2 Data Formats 6
2.3 Level 1 Processing 7
2.4 Level 2 Processing 7
2.5 Level 3 Processing 8
2.6 Level 4 Processing 9
2.6.1 Static Catalogs 10
2.6.2 Forced-Photometry Catalogs 11
2.6.3 Difference-Imaging Catalogs 11
2.7 Level 5 Processing: Community-Contributed Products 12
3 Supporting Tools and Services 12
3.1 Empirical Point-Spread Functions 12
3.2 Idealized Simulations 13
3.3 Instrument-Signature Simulations 13
4 Data Analysis Software Infrastructure 14
4.1 Development schedule 15
5 Data Access 16
6 The Roman Science Platform 17
7 Services and Tools that are not currently in-scope for the SOC 18
8 References 19

1 Introduction

The Roman WFI instrument consists of 18 near-infrared detectors, each with 4096x4096 pixels. Telemetry data from the instrument will arrive at the SOC, where they will be processed into data products that are suitable for scientific analysis. This document outlines the data processing steps that are currently planned at the SOC and describes the resulting data products. Details are subject to change as the development progresses. Some of the data processing steps (for spectroscopy and for the bulge time-domain survey) are carried out at the Roman Science Support Center at IPAC, as illustrated in Figure 1.

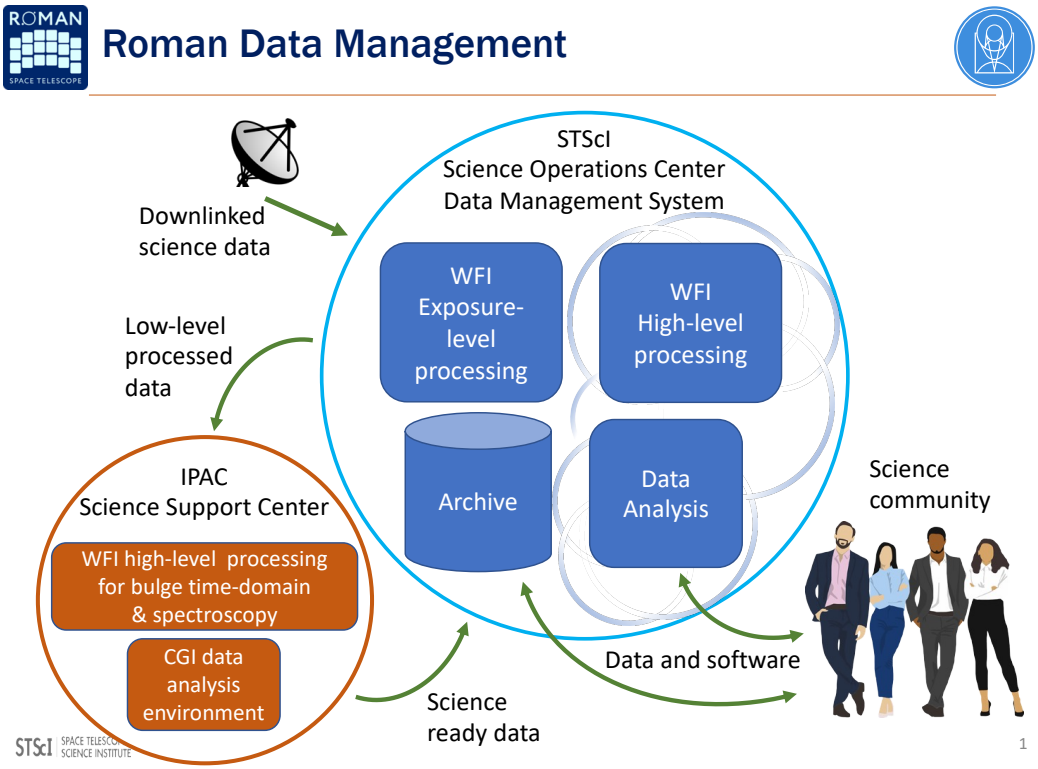

Figure 1: Data Management for the Roman Wide Field Instrument. Basic processing through removal of instrument signatures is done at the SOC. The SSC is responsible for higher-level processing of the spectroscopy and bulge-time-domain data. The SOC handles data high-level processing for the rest of the WFI data (as described later in this document). All of the data comes back to the Roman Science Archive hosted by the SOC. Some of the data processing and archive holdings are actually in the Cloud (at AWS), although for the most part this is transparent to the users. There will be a Science Platform in the cloud where users can access the Roman data with high-bandwidth connections to the most commonly used data products.

Science data from the Roman mission have no proprietary period and are made immediately public. Processed data will be made available on timescales that range from 48 hours (for the individual calibrated "Level 2" images) to 6 months (for uniformly processed survey data). Further details are provided in later sections.

For many science applications, the volume of WFI data needed to carry out the investigation will exceed what can be easily transferred to or processed on an average workstation. The SOC will provide computational support, and high-bandwidth access to the Archive cloud data stores, through the Roman Science Platform. The computational resources will be scaleable to support

large amounts of data processing.

The rest of the document describes these aspects of the SOC support in more detail. Section 2 describes the WFI pipeline and data products. Section 3 describes supporting tools that are being developed or customized for Roman, including those to provide high-quality point-spread functions (PSFs), and those to facilitate simulating Roman data. Section 4 reviews the data-analysis tools that are likely to be available in the Python astronomy ecosystem, many of which have been developed and are being maintained for the *Hubble* Space *Telescope* (HST) and the *James Webb Space Telescope* (JWST). The Roman Project is not currently planning to develop new data analysis tools for use with Roman data but will maintain currently available tools and ensure compatibility with Roman data. Section 5 summarizes plans for the Roman data archive and outlines the tools and data-access methods that are expected to be available. Section 6 describes the plans for the Roman Science Platform, a cloud-based computing platform that will be available to the community for analyzing Roman science data. This platform offers scaleable computation and high-bandwith access to the most-commonly used data products. Finally, section 7 highlights some of the services and tools that are not currently in the funded baseline for the SOC.

The data-processing plan has evolved and will continue to evolve in consultation with the astronomical community. For the past several years, most of this interaction has been through the Science Investigation Teams, with regular meetings to discuss detailed topics of algorithms and software. Regular technical exchanges with the Roman science community will continue as development progresses.

2 Pipeline Data Products

The Roman science data calibration pipeline software will be based on the calibration software developed for the JWST but adapted as appropriate for use with Roman WFI detector data. The science data calibration pipeline takes science data in a standard archival format and mathematically applies calibration algorithms and calibration reference data to remove instrumental and environmental signatures from the data. This results in science data products that are prepared for further scientific investigations. In later stages, groups of exposures are combined and higher level information is extracted (e.g. source photometry).

The SOC WFI science data reduction pipeline will process WFI exposures and groups of exposures, into various levels of pipeline data products, as outlined in Table 1. At the single exposure level, detector and instrument effects are removed from Level 1 data products (uncalibrated files) to produce Level 2 data products (calibrated detector files). In addition to exposure level processing, DMS will also perform higher level data processing that may combine data from multiple exposures to produce mosaics (Level 3 data products) and catalogs of information (Level 4 data products).

The `romancal` software package enables processing of data from all WFI detectors and observing modes to Level 2 data, producing fully calibrated individual exposures. At this point, the processing responsibility splits and the SOC will only create higher level data products from

WFI imaging mode observations. The SSC pipelines will accept the SOC Level 2 spectroscopic mode products, create higher level data (Level 4) and return the resulting products to the SOC Archive. SSC will also create and return higher level data products for the Galactic Bulge Time Domain Survey. The algorithms and overall structure of the pipeline are approved by an inclusive working group that regularly consults the appropriate Roman calibration working groups. As is the case for JWST, the `romancal` software package, including details about how to install and run the calibration pipeline, will be publicly available. Each defined step in the pipeline can be run with its own configuration file specifying the parameters to be used for that step.

Table 1. Levels for the Roman WFI pipeline data products.

Data Level	Description	Comment
0	**Packetized data as delivered from the spacecraft**	Level 0 data is raw packetized science telemetry as transmitted from the Observatory and received at the ground stations.
1	**Uncalibrated "raw" individual Exposures**	Level 1 products are uncalibrated individual exposures for each detector consisting of raw pixel information formatted into the shape of the detector.
2	**Calibrated Individual Exposures**	Level 2 data products are corrected for instrument artifacts, with slope fitting, outlier rejection, and other procedures to obtain a true mapping of the scene flux. Calibrated exposures have appropriate astrometric and geometric distortion information attached, and with the exception of grism/prism data, are in units that have known scaling with flux. Uncertainty and data quality arrays are provided.
3	**Data Resampled to a Regularized Grid and Combined**	Level 3 products are groups of calibrated exposures resampled to a regularized grid, removing the geometric distortion of the original pixels. Relevant images taken in the same filter are co-added. Uncertainty arrays are provided.
4	**Derived Data**	Level 4 products are usually focused on sources/objects rather than pixels or celestial coordinates. These can contain traditional data (such as positional, size and shape information) or complex data such as extracted spectra or postage-stamp images of the relevant source from all contributing images.
5	**Community-Contributed Products**	Community generated data products that can be of arbitrary form and complexity. These encompass any data that is returned to the SOC for archival storage by contributing scientists or groups and may include data that could be described as belonging to any of the previous data levels.

The `romancal` software package is written primarily in Python. It will be included with the Roman Science Platform, and easily installable by users on compatible workstations and laptops. The parameters of the various pipeline steps are configured in easily editable text files. This enables a wide variety of custom processing without the need to write new code. Typical use cases may involve tweaking the parameters for cosmic-ray rejection or the pixel scale (and hence oversampling of the point-spread function) in co-added images. The pipeline code itself is modular and has been written with the aim of making it relatively straightforward to modify to add new steps or change algorithms.

Level 2 data will be processed and made available through MAST within 48 hours of receipt, by the SOC, of all data necessary to complete processing. Level 3 science data will be available within 5 days of receipt of all the needed products

2.1 Survey Data Releases

While the standard processing makes fully pipeline-processed data available within the few-day timescales mentioned above, the most useful SOC data products may well be the more-uniform survey data releases.

All Level 3 and Level 4 science data – processed (or re-processed if necessary) with the most appropriate calibration data – will be available 6 months after the last relevant data set is obtained. This does not necessarily imply a 6-month cadence for the data releases. The timing and processing details of the data releases will depend on the scheduling of the observations and on the observing strategies that are ultimately adopted for the surveys. (For example, the pixel scale of the rectified co-added Level 3 images could in principle be optimized based on the details of the dithering strategy.) The primary goal is for these data releases to be the definitive data set in terms of the instrument-signature removal and relative astrometry. There will be a set of co-added Level 3 images and Level 4 catalogs, aimed at serving a broad range of science. It is impossible to optimize these data products for all types of science, so it is likely that the community will also produce different versions of co-added images and catalogs.

In the course of achieving the exquisite calibration accuracies required for realizing the Roman core science goals, the science teams are expected to provide refinements in photometric calibration and astrometric accuracy as well documenting and possibly correcting for systematic effects that influence cosmic shear measurements. Catalogs constructed by the science teams can take advantage of survey-level self-calibration or cross-calibrations with other data sets. The pipelines can be optimized for specific science purposes. The final versions of these catalogs will be hosted in the Roman Science Archive as Level 5 products. These refined catalogs are expected to meet the mission's core science requirements, whereas the SOC catalogs are not specifically tailored for that. The SOC will work with the science teams to ensure that appropriate data-models and schemas for their products are created and that submitted data products conform to expectations.

2.2 Data Formats

The Level 1-4 science data products will be created using the Advanced Scientific Data Format

(ASDF). This data format is also being used by JWST. Either pure ASDF files or FITS files can be used as the basis for JWST calibration reference data. In the case of JWST science data products, which are saved as FITS files, the ASDF-formatted metadata are stored in a FITS extension. The JWST and Roman data-processing software works with data models that are independent of the storage format. Conversion to and from FITS or ASDF is confined to input/output routines, making the storage format transparent to the user for many purposes. For those needing to work directly with the files, the ASDF serialization is fully specified in schemas that are independent of any particular programming language.
It will be possible to obtain FITS-formatted Roman data files from the SOC Archive that contain the basic required FITS header keywords and data arrays, while storing the complete set of metadata in an ASDF extension (referred to as ASDF-in-FITS). For both JWST and Roman, the full WCS, which includes the geometric distortion models, are stored as part of the more complex ASDF metadata (which are stored in a FITS extension in the case of JWST but will be in the native ASDF file for Roman). Users that need direct access to this information at lower data levels (Level 2) will need to use the software tools provided by the SOC or develop additional tools that can understand the complex models that ASDF naturally allows for metadata storage. The SOC will provide complete descriptions of the Roman data models and their contents in order to support science teams who wish to write their own tools to read the serialized ASDF files directly. An example would be a Fortran interpreter for the ASDF data format. For higher level data (Level 3), after the distortion has been removed from imaging mode exposures, the simpler FITS WCS representation can be used and saved with the requested FITS file.

2.3 Level 1 Processing

In the first stage of processing, the Roman DMS pipeline gathers science and engineering telemetry to create a single file for each detector and each exposure. The individual non-destructive readouts of the WFI H4RG detectors may or may not be combined before being transmitted to the ground, the combination is controlled by predefined exposure specifications. Each Level 1 data file that is created formats pixels from the downlinked reads into a single file that contains 3D data arrays. These arrays are the full size of each detector, and contain all downlinked reads. The resulting data product is ready for input into the exposure level calibration software.

2.4 Level 2 Processing

Level 2 processing applies calibration steps to remove the instrument signatures from the data and combines the information from the individual (or groups of) readouts. Figure 2 shows a diagram of the anticipated flow. Most of these steps are standard for near-infrared detector processing, including for the detectors on JWST. These are described in detail in the JWST documentation at jwst-docs.stsci.edu. The greyed-out ovals are steps that have been considered but are likely not to be performed for the Roman WFI data. Most of the Roman WFI data will be bias-subtracted onboard, so the bias correction may not be needed in the pipeline.

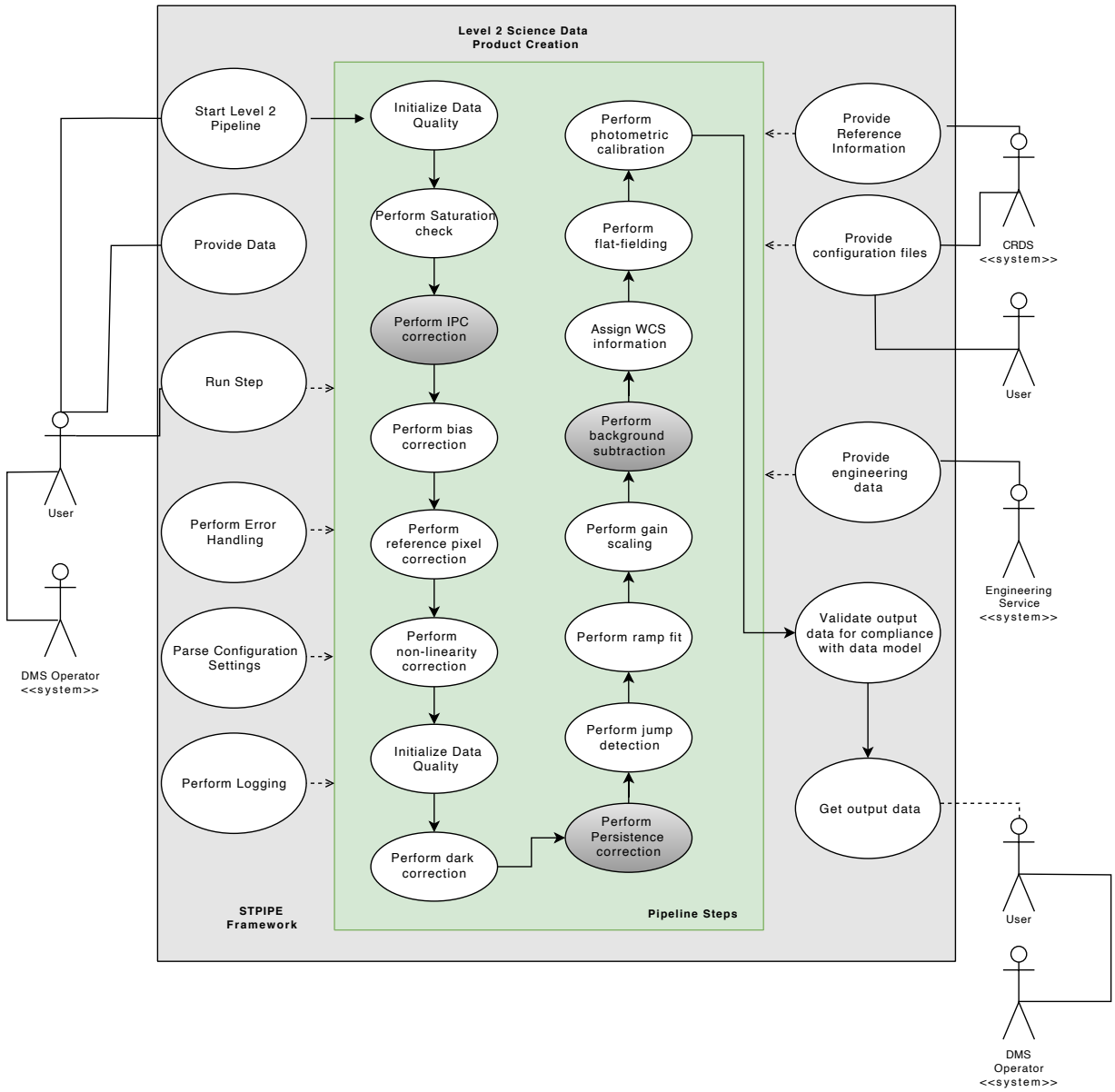

Figure 2. Processing steps involved in producing Roman WFI Level 2 data products.The grey ovals are steps that have been considered but are currently not planned to be executed for the Roman WFI data.

2.5 Level 3 Processing

Level 3 processing resamples and projects the data onto a tangent-plane (removing the geometric distortions from the telescope). The pipeline will co-add relevant datasets taken through the same filter on overlapping patches of the sky which have been explicitly associated with each other. Before doing the coaddition, the pipeline will refine image alignments, match sky backgrounds, and mask outliers such as cosmic rays. The details of how the sky will be tessellated in the resulting co-added images remain to be finalized. Figure 3 summarizes the steps in Level 3

processing. The drizzle resampling algorithm (Fruchter and Hook 1998) is the method currently envisioned for use in rectifying the data. The implementation used with the Webb calibration pipeline software will also be used with Roman. Level 3 data products will be available from the SOC Archive within 5 days of receipt of the last data needed to create them. The data-release versions of these, (re)processed with consistent calibration across the full survey area, will be available 6 months after the receipt of the last relevant survey data.

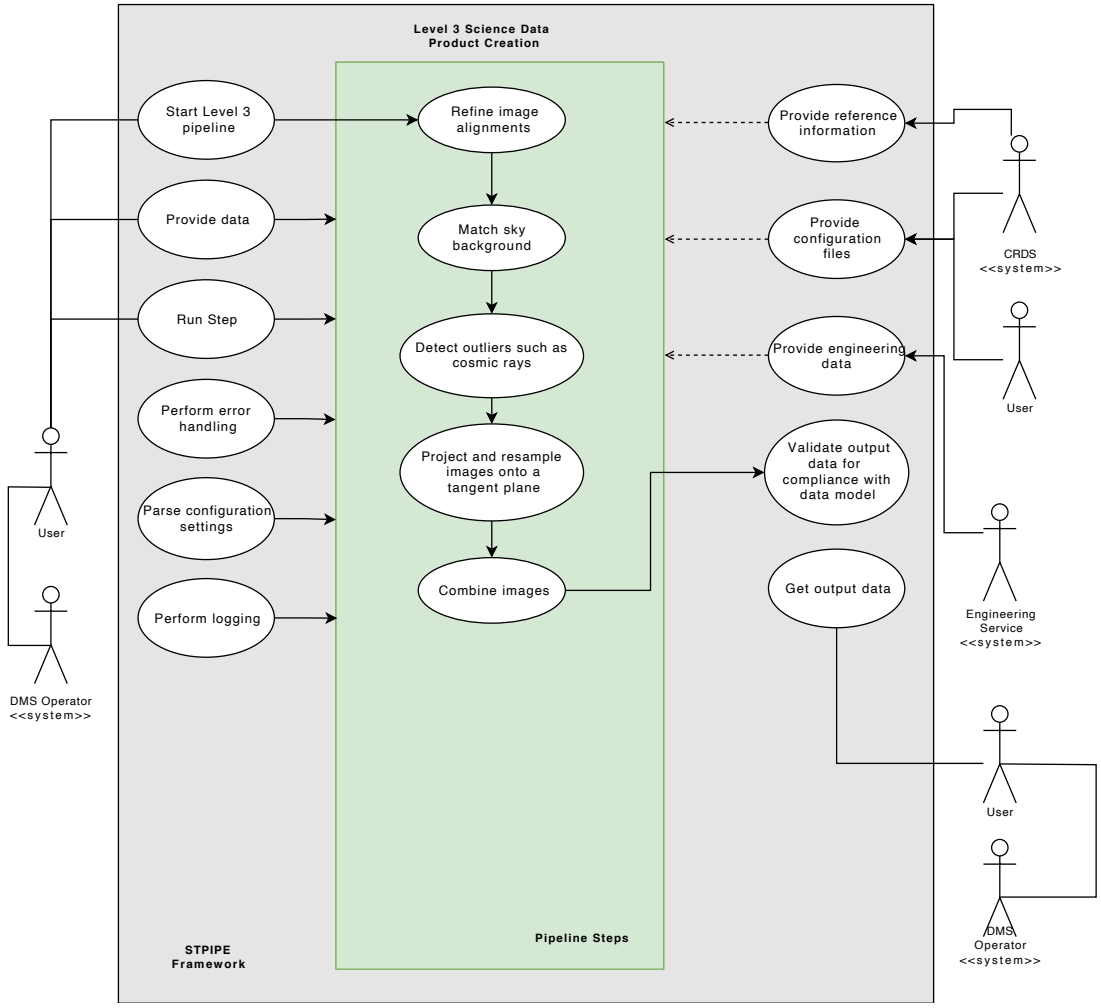

Figure 3: Steps involved in producing the Level 3 data products from the Level 2 input images.

2.6 Level 4 Processing

Level 4 data products consist of catalogs and other derived or associated information. Many of the details are still to be decided based on science-community input. There will be several different types of catalogs:

1. Catalogs associated with each Level 3 image (whether individual exposure or co-added). These will be available within 7 days after the the last data needed to created them is received by the SOC.

2. Catalogs associated with the co-added Level 3 images, will be available with each data release. Data releases are anticipated to be roughly 6 months after the last relevant data are obtained. These astronomical source catalogs will be available for query through MAST.
3. Forced-photometry variability catalogs. These are intended to find both objects that vary in flux and those that vary in position between observations. These catalogs will be available on the 6-month data-release timescale.
4. Difference-imaging catalogs for identifying transients. These are nominally available with the survey data releases, but having them available on the 48-hour timescale of the initial data processing is a possibility under discussion. The template images will be constructed from all prior data taken using the same filter, on the same patch of sky, as the new image.

2.6.1 Static Catalogs

The catalogs of the "static sky" (items 1 & 2 in the list above) will follow the standard procedures used for making object catalogs (e.g. using the SExtractor or Photutils packages). The key steps included are the following:

- Estimate and subtract background
- Convolve with a detection kernel
- Identify connected pixels above a noise-dependent threshold
- Hierarchically de-blend overlapping sources
- Measure fluxes through apertures (with and without convolution by a PSF-matching kernel to correct all photometry to one reference PSF)
- Measure shapes
- Classify stars vs. galaxies based on shapes (only)
- Compute photometric redshifts from multi-band (Roman only) photometry

The following outputs will be available:

- Level 4 catalogs including uncertainties computed from a noise model
- Photometry, positions, shapes, and local background estimates
- Level 4 segmentation maps
- Catalogs of input & output parameters for injected artificial sources (for a relatively limited variety of artificial source parameters and iterations). More details on source injection are in Sections 2.6.1.1 and 3.2.

2.6.1.1 Source injection

The catalogs described above will be accompanied by the outputs of a source-injection pipeline that will insert synthetic sources into the WFI data, and catalog those sources following the same procedures used for real sources. The details of these simulations have yet to be decided, but are likely to cover only the most obvious part of the parameter space: e.g. point sources and Sérsic-profile galaxies with a range of ellipticities, Sérsic indices and half-light radii. It is likely that the number of simulated sources in these analyses will contain less than 1% of the number of real sources. The source-injection pipeline will be available to the community for re-use and

modification.

2.6.2 Forced-Photometry Catalogs

The strategy for constructing a variability catalog is under discussion. The strategy involves using object catalogs constructed from the co-added Level 3 data associated with the survey data releases as the input. Steps include:

- Identify point-like objects
- Identify relevant Level 2 images
- Transform coordinates of objects from RA and Dec. to Level 2 pixels
- Measure photometry without re-centering the source
- Estimate the position of the source and determine if it might have shifted
- Flag possible moving sources
- For moving sources, measure photometry at the shifted position
- Apply photometric corrections based on PSF and jitter information
- Compile the time series
- Compute variability indices
- Flag data-quality issues that might be uncovered in the time-series data

There are no current plans to complement this catalog with the results of artificial source injection.

2.6.3 Difference-Imaging Catalogs

A catalog based on difference imaging is useful for identifying variable sources and transients that might not appear in the co-added images, or might be superimposed on a host galaxy. The following briefly summarizes the strategy under consideration for Roman. As each new Level 3 image is constructed:

- Identify overlapping Level 3 images through the same filter that have already been processed.
- Co-add these images to make a template image.
- Convolve with a PSF-matching kernel if needed to match the new observation.
- Subtract the template from the new image.
- Identify point sources that exceed a threshold.

The output is a catalog associated with each individual Level 3 image that lists fluxes and positions and associated metadata and data-quality information for each identified transient.

There are no current plans for the following:

- consolidating these files into a time-series or cross-matching them with the static catalogs from the survey data releases;
- complementing this catalog with the results of artificial source injection;
- providing alerts. - (While there will not be transient alerts, there will be a subscription service for notification when user-selected new data products appear in the archive.)

2.7 Level 5 Processing: Community-Contributed Products

Level 5 data sets are those that are generated by the science community and provided back to MAST for archiving and distribution. Examples might include photometric redshifts based on complementary data sets from other facilities, value-added catalogs of derived properties (e.g. star-formation histories from fitting spectral-energy distributions), catalogs that contain image shapes or revised photometry based on survey-level calibration of PSF effects, or selection functions for a particular class of objects based on extensive artificial source injection. The details and cadence of such data releases have yet to be defined and will be generally up to the teams responsible for creating these data products. However, requirements for data formatting, content, and acceptance will be documented and made clear prior to submission.

It is anticipated that science teams will create catalogs that also include survey-level calibrations (e.g., ubercal, cross-calibrations with other surveys, manual optimization of pipeline parameters for specific science purposes) that go beyond what the SOC catalogs provide. These will be hosted at the SOC as Level 5 data products. The final versions of these core-survey catalogs are expected to meet the mission science requirements, whereas the SOC catalogs are not specifically tailored to that. Data processing to meet the core science goals is also expected to benefit from being tailored to the specific science goals and observing strategy, while the SOC pipeline is intended to be more general. The SOC will work with those science teams to ensure that appropriate data-models and schemas for their products are created and that submitted data products conform to expectations.

3 Supporting Tools and Services

Several important tools and services will assist in ensuring high-quality science-ready data from the SOC pipelines. They are the following:

1. A curated library of high-quality empirical point-spread functions (PSFs)
2. Simulation tools for creating synthetic Level 1 images in the appropriate format with some of the most important instrument signatures included.
3. Simulation tools for creating idealized synthetic Level 2 and Level 3 images – with realistic PSFs, geometry, and noise, but without the other instrument signatures.

These tools go beyond what has been provided for Hubble and Webb. They are necessitated by the need to reduce the duplication of effort by the community, given the survey data volumes. The relatively homogenous and predictably known nature of the Roman WFI observations also make providing these tools and services more feasible than they have been for Hubble and Webb, which have a wide variety of observing modes and diverse observing strategies.

3.1 Empirical Point-Spread Functions

While Roman is designed to provide very stable point-spread functions, measuring precise PSFs is crucial to accomplishing the scientific objectives of the mission. The WebbPSF modeling tool is available now for modeling Roman point-spread functions. However models need to be tested with empirical data. Furthermore, the WebbPSF models do not include the effects of the detector pixels. A complementary approach is to construct empirical PSFs (ePSFs) using dithered

observations of point sources (Anderson & King 2000). The current plan for Roman is for the SOC to construct an extensive ePSF library that will be used both by the pipeline in making astrometric corrections and in making catalogs, and will be available to the community for performing custom data analysis. The inputs for generating the ePSFs are individual dithered Level 2 images. These dithered observations are used to construct a grid of high signal-to-noise-ratio, oversampled, ePSFs that span the entire focal plane. These will be stored in a database with accompanying metadata. The most obvious metadata items are position, filter, and time. However, as the mission progresses it may be possible to identify engineering parameters (temperatures or jitter) that can also be useful for selecting the appropriate ePSF for modeling a particular observation.

The ePSF database will have an Application Programming Interface (API) that allows the SOC pipelines and science users to select the appropriate ePSFs. Interpolation between ePSFs and jitter can then be added as necessary by the user or downstream software.

3.2 Idealized Simulations

Simulations that either construct full images of synthetic sources or insert them into actual Roman images are essential for achieving many of the Roman science goals. The SOC will be providing tools for making such simulations (building on existing tools), and will be doing some limited artificial source injection as part of making the data-release catalogs.

The inputs to such simulations are fluxes and spectral-energy distributions (SEDs) of individual sources and parameters that describe collections of sources (e.g. luminosity functions and size distributions). The simulation tools will perform the following operations:

- Apply the instrumental throughputs (using synphot-like software) to convert the SEDs from physical units into counts per second
- Convert the model shapes into an image
- Convolve with the point-spread function
- Add background
- Apply geometric distortions (if simulating Level 2 outputs)
- Resample to the appropriate detector pixel geometry
- Add noise

The images will be in the appropriate data format for Level 2 or Level 3 Roman images so that they can be fed through subsequent stages of the science pipeline.

It will be possible to use these tools modularly – for example to provide an image data cube and convert this into appropriately-sampled Roman-like images with noise, rather than using the SOC-provided tools to make this data cube.

3.3 Instrument-Signature Simulations

Simulating data with the most important instrument signatures is going to be essential for validating the SOC calibration pipeline. The tools for making these simulations will also be

useful to the community in planning observing and data-analysis strategies. However, it is likely to be too costly computationally to run source-injection simulations from this stage all the way through the pipeline on the scale needed for constructing selection functions. Hence the current plan is not to optimize these simulations for speed or convenience in a pipeline. The inputs to these simulations are count-rate images using the native (distorted) pixel scale that have already been convolved with a PSF but have no noise. The tools will simulate the downlinked reads which may include reads which are the combination of multiple individual readouts (as specified by the predefined exposure information). The following instrument signatures are currently planned to be included:

- Cosmic rays
- Nonlinearity
- Interpixel capacitance (if not using an empirical PSF)
- Basic persistence model
- A simple model of readout noise and dark current (details still to be decided)

The following features are currently not planned to be included:

- Intra-pixel sensitivity variations
- Brighter-fatter effect
- Sophisticated readout model (e.g. $1/f$ bias drifts)

The simulator software will be modular, making it relatively straightforward to modify the treatment of individual instrumental effects or add additional ones. The output images are in a format that can be fed directly into the Level 2 processing in the SOC calibration pipeline.

4 Data Analysis Software Infrastructure

Beyond the Roman-specific tools mentioned in section 3, there is an evolving ecosystem of tools to support data analysis from Hubble, Webb and other facilities. These are open-source tools, but a significant fraction of the financial support for their development has come from NASA. The roadmap for continued maintenance of these tools for the duration of the Roman mission is unclear. Nevertheless it is worth summarizing a few of the more prominent tools that are currently in widespread use and provide the kinds of functionality that will be needed for Roman data analysis. STScI plans to continue to distribute these tools until at least the end of the Webb mission. An essential underpinning of our pipeline software development and analysis tools is the suite of libraries on which they are based. These libraries also enable significant development of analysis tools by astronomers themselves.

The suite of post-pipeline data analysis tools is intended to help astronomers with the often iterative and interactive workflow involved in converting these science calibration pipeline data products into meaningful scientific results. This involves tasks such as:

- inspecting data and data quality information;
- masking or flagging data and using those annotations to guide later steps in the analysis;
- using the results of interactive analysis to guide a custom run of the pipeline (e.g., tweaking spectral extraction parameters or background estimates);
- performing optimized spectral extraction techniques;

- combining data sets in various ways, with careful attention to astrometry, PSF matching, and other issues;
- source detection and photometry using different choices or algorithms than those used in the pipeline;
- measuring lines and continuum in spectral data;
- fitting models to data or otherwise testing hypotheses.

A typical workflow involves highly interactive exploratory analysis on small portions of the data, followed by the development of custom scripts to automate the analysis on larger data sets. Further details and demos are available in the JWebbinar series. The table below lists some of the tools and libraries most likely to be relevant to Roman.

Table 2. Tools and libraries likely to be relevant for Roman data analysis.

Tool	Purpose
`Astropy`	A general-purpose python library for astronomy. Modules include support for common data formats, constants and units, cosmology, fitting, filtering, time and coordinate systems and statistics.
`glue`	A python library to explore relationships within and among related data sets
`photutils`	Tools for detecting and performing photometry on astronomical sources
`specutils`	Tools for spectroscopic analysis
`jdaviz`	Interactive analysis tools. The Specviz, and Mosviz modules are particularly relevant for Roman Specviz is a 1D spectra analysis tool with similar functionality to IRAF's splot. Mosviz is an analysis and visualization tool for multi-object spectroscopy
`asdf`	Advanced Scientific Data Format is an interchange format for scientific data
`gwcs`	Generalized World Coordinate System tools for dealing with image and spectral geometries. These geometries can be serialized in ASDF files and are supported by the rest of the astropy ecosystem.
`synphot`	Synthetic photometry toolkit for building model spectra and estimating count-rates.
notebooks	Curated and maintained interactive Jupyter notebooks that illustrate common steps in data analysis using the tools in this ecosystem. These can be used as starting points for customizing a data-analysis workflow.

4.1 Development schedule

Development of the software to support Roman is already well underway. Most of the science software development is done on Github, and includes continuous integration and continuous development tools. Thus, while some of the packages, such as `romancal`, can be downloaded even now, there will be significant evolution of functionality as-development progress. To aid in planning, Table 3 provides a rough sequence of events. The years 2022 and 2023 are the most important for interaction with the community on fundamental decisions concerning measurements and algorithms. As we get closer to launch it becomes progressively harder to make major changes to the operational pipeline. While there will undoubtedly be revisions during the mission itself, any major change comes with a cost of creating non-uniformity in the

data products and/or requiring reprocessing of large volumes of data.

Table 3. Roman WFI data processing software development schedule.

Year	Highlights
2022-23	• Pipeline definitions and data product information • Calibration Reference Data System (based on Webb's) • Photometric calibration pipeline steps (relative and absolute) • Mature software: Webb pipeline, WebbPSF, Pandeia, STIPS, Webb data analysis tools • Trades and Algorithm development, with significant community input: Simulations, PSFs, time-domain processing, astrometry, object shape measurements, survey mosaic geometries, relative and systematic uncertainties, science platform
2024	• Roman Science Platform • Archive Data Access Application Programming Interfaces (DAAPIs) • Pipeline procedures for precise alignment and mosaicking of images • Ability to ingest Level 5 products from the community • Cross-matching tools in the archive • Instrument-signature simulation tools • Idealized simulations and source-injection simulation tools
2025	• Pipeline for static-sky catalogs including matched-PSF photometry • Point-spread function library infrastructure
2026	• Absolute astrometric calibration pipeline
2027	• Difference imaging • Populated empirical PSF library

5 Data Access

The SOC Roman Archive holds all the Roman data that scientists need to achieve the mission's scientific objectives. This includes all levels of science data products from the WFI and CGI instruments. It also includes engineering data, calibration reference files, and ancillary data that are critical for the processing of scientific data. Community-contributed products delivered to the SOC are also available.

The Roman Archive holdings will be distributed between the cloud (Amazon Web Services; AWS) and storage on premises at STScI. Current plans store Level 1 data on-premises at STScI, Level 2 data both on premises and in the Archive cloud holdings, Level 3 and Level 4 data files in the Archive cloud holdings, Level 5 products on premises, and all MAST databases on premises. The details of which data sets go where may evolve as cloud services and costs evolve. Nevertheless, the location of the data will remain transparent to the archive user and the fastest and most cost-efficient delivery source will be used to provide the data to the user when they are requested.

The MAST archive for Roman will have a web portal with an array of services that will be familiar to Webb users. The archive will provide query forms that facilitate searching for datasets or for selecting sources within catalogs. There will be interactive tools for visualizing both imaging and spectroscopic data.

Because the Roman data volume is enormous, it is likely that there will be some monitoring and throttling of requests to manage overall costs. There will be ways to accommodate users with substantial needs, but the system is not being designed with the capacity to support an annual egress data volume that is much more than about 30 times the annual volume of data being generated by the pipeline.

6 The Roman Science Platform

The enormous data volume necessitates different ways of thinking about data analysis. A typical astronomer's desktop computer will not have the capacity to store and analyze an entire Roman core community survey. The data volumes involved also exceed the capacity of most departmental servers for storage and computing. It makes more sense to move the computing to the data rather than the data to the computing.

To accomplish this, the Roman Science Platform (RSP) will be deployed and maintained in the commercial cloud at AWS. The data will be available at high bandwidth from AWS S3 storage. The amount of computation is easily scaleable, so it can grow and shrink in response to community needs.

The RSP will use JupyterLab and JupyterHub to provide a familiar collaborative computing environment for the science community. These tools are already being implemented and explored with Webb science data products and the Time Series Integrated Knowledge Engine (TIKE) that is employed by MAST. The full suite of Roman science calibration pipeline and analysis tools will be installed and maintained on the platform. Users will be also able to install their own RSP compatible software and run it from a Unix command line, provided through the terminal available with JupyterLab, or from Jupyter notebooks.

The full details of the platform services have yet to be finalized – in consultation with the Roman science community, and with the benefit of experiences with existing science platforms. The concept is for the computing instance to be running a standard version of a Linux operating system. Users will have access to an up-to-date Python environment with the most common scientific libraries already installed. There will compilers available for other common scientific programming languages such as C, C++ and Fortran. Scientists using the platform will be encouraged to use Github to develop and share software. Integration between JupyterLab and Github is already quite mature. STScI maintains a set of public style guides to encourage coding best practices.

Several different tiers of platform access are anticipated, with justifications needed for obtaining access to the tiers with more resources. Investigator grants will not be charged individually for the first three tiers. All tiers require users to have a MyST account. The tiers are conceived to be as follows:

- Entry Tier
  - Suitable for exploring Roman data or processing a few square degrees.
  - Available to all users with a MyST account.
- Research Tier
  - Suitable for conducting research on the scale of a typical publication. Typical

analysis of tens to hundreds of square degrees.
    - Can be requested with a brief justification via a web form.
    - Research Tier Allocations (and/or Team Tier) will be automatic for those with approved NASA Roman grants.
- Team Tier
    - Allows pooling of resources within a team.
    - Sufficient allocations for custom processing of hundreds to thousands of square degrees.
    - Requires more substantial justification, usually as part of a NASA grant proposal, but there will be a process to request Team Tier allocations even without a NASA grant.
- Special cases
    - It is possible to create larger allocations than anticipated for the team tier.
    - Conceptually, these will be requested as part of future Roman General Investigator or NASA/ROSES proposal processes.
- Access from other AWS accounts
    - In some cases it may be more practical for teams to have their own AWS accounts. They will have access to all the SOC science data processing software and access all of the Roman data. However, they will have to manage users and resources on their private platform using their own resources, rather than benefiting from the SOC platform management infrastructure.

7 Services and Tools that are not currently in-scope for the SOC

Design of the Data Management System has included many trades and compromises. The following are some of the scientifically important items that are currently beyond the scope of the products and services to be provided by the SOC. In some cases, the work is strongly dependent on a particular survey strategy, or depends on expertise or simulations relevant to specific science goals. In other cases, the work would extend the scope of the SOC activity beyond what current budgets allow.

- Catalogs with survey-level cross calibration, specifically tuned to the science needs of weak lensing or supernova cosmology.
- Scientific validation of the core science goals.
- Development of specialized algorithms optimized for relatively narrow scientific objectives. Examples might be motivated by specific mission science requirements (e.g. special preprocessing of the images prior to measuring galaxy shapes), while others might serve science objectives outside of the Roman core areas.
- Joint processing of complementary datasets from other facilities (e.g. Rubin or Euclid).
- Storing complementary datasets from other facilities on AWS for access from the Roman Science Platform.
- Using data from other facilities in computing photometric redshifts.
- Transient alerts.
- Development of new visualization tools optimized for the large Roman data volumes.
- Development of new general-purpose data-analysis tools.

Many of these items are either necessary to meet the Roman core science goals and success criteria, or have been identified by community reviews as important scientific uses of Roman data. The science teams needing specialized processing may be the best suited for developing the corresponding software, services or tools. Thus, it is the expectation that many of these items and/or additional tools or processing complementary to the SOC baseline will be proposed and selected as part of NASA/ROSES or Roman General Investigator opportunities.

8 References

Anderson, J. and King, I. R., 2000, PASP, 112, 1360.