STScI | SPACE TELESCOPE SCIENCE INSTITUTE

EXPANDING THE FRONTIERS OF SPACE ASTRONOMY

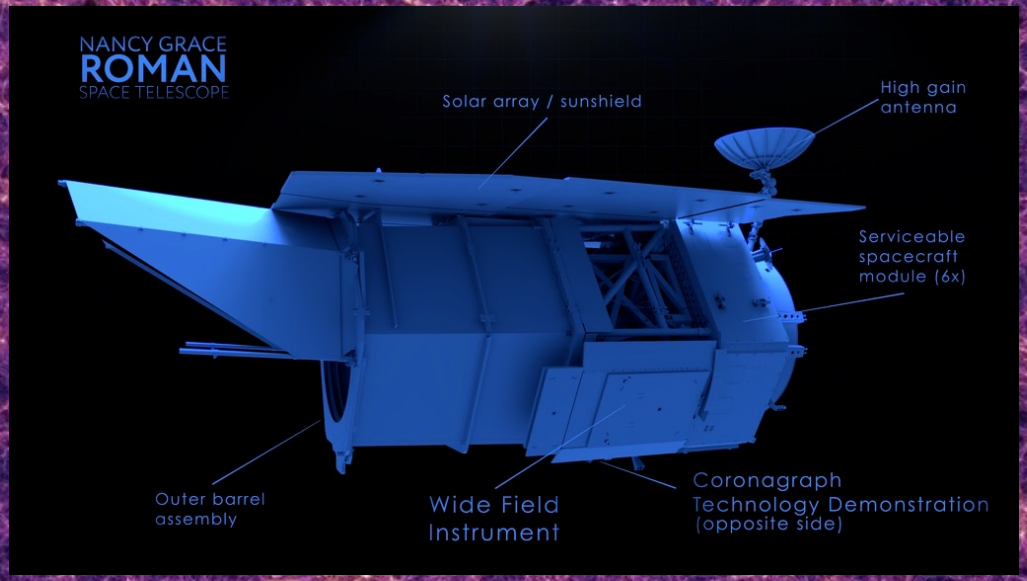
The Nancy Grace Roman Space Telescope Data Management Challenges & Solutions

Harry Ferguson
STScI

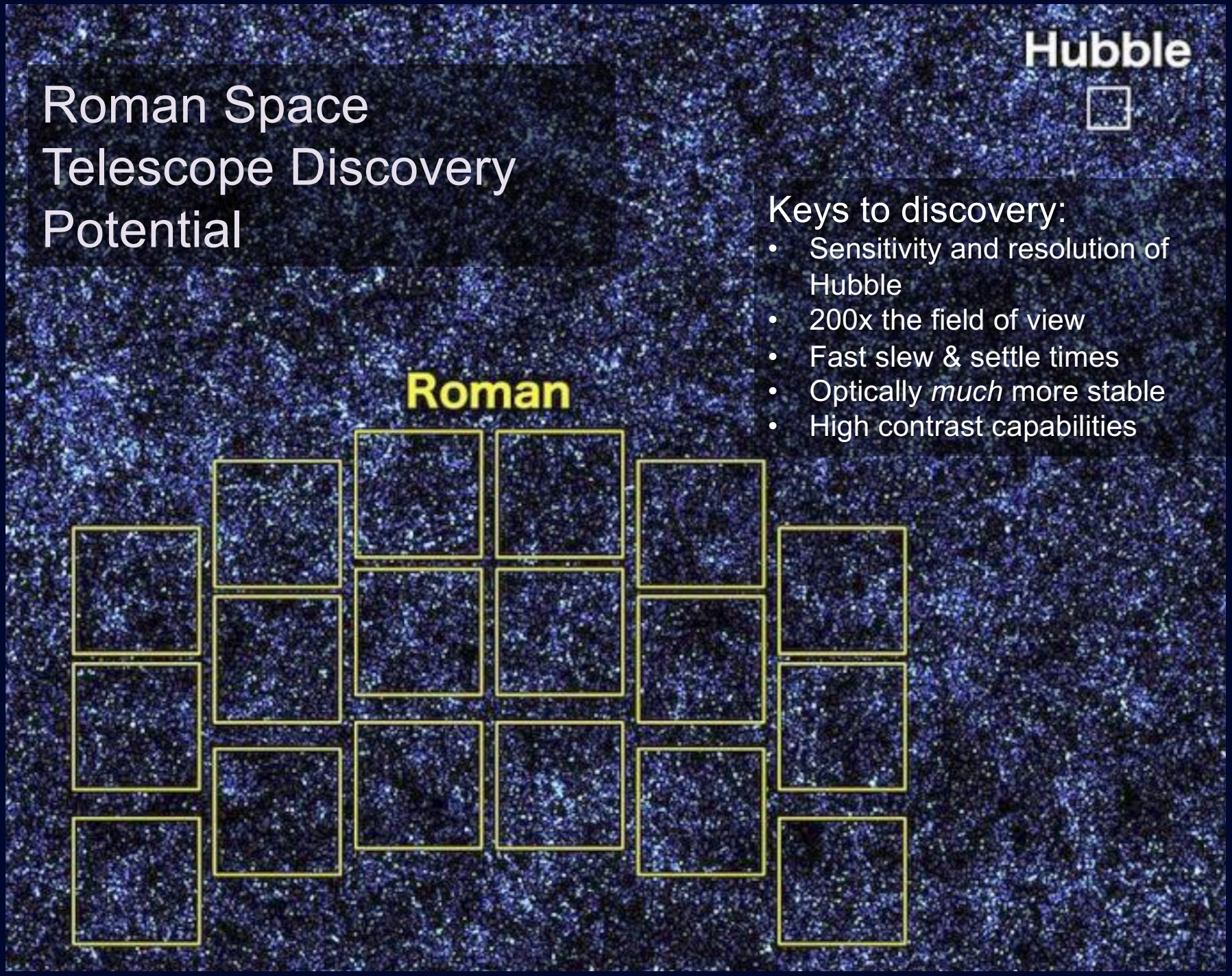
June 3, 2025

Launch 2026

2.4 m primary mirror
0.28 sq. degree field of view
0.5-2.3 μm



Roman Space Telescope Discovery Potential



Keys to discovery:

- Sensitivity and resolution of Hubble
- 200x the field of view
- Fast slew & settle times
- Optically *much* more stable
- High contrast capabilities

Core Science

- Cosmology: Three independent probes of dark energy
  - Baryon Acoustic Oscillations (geometry)
  - Supernova standard candles (geometry)
  - Weak Lensing (geometry and growth)
- Exoplanets:
  - Microlensing bulge survey (time-series imaging)
  - Coronography technology demonstration
- General Astrophysics: (at least 25% of prime mission)
  - Everything else, including…
    - Resolved stellar populations in the nearby universe
    - Stellar motions and distances
    - Galaxy evolution, deep fields
    - Milky Way structure & evolution
    - Small bodies in the solar system

Science Operations and the Data Management System

STScI Roman Science Operation Center responsibilities include:

- Planning & scheduling all observations
- Calibration and support of the Wide Field imaging
- The archive (MAST) for all mission data
  - Most Roman science will be archival due to the survey nature of the mission

NASA Astrophysics *Big Data*:

- Data accumulated per week likely to be >>100x *Hubble*
- Both catalogs and pixel-level data sets provide unique science opportunities
- Downloading and processing exceeds resources typically available

Barbara A. MIKULSKI ARCHIVE FOR SPACE TELESCOPES

Science data products from multiple mission partners

- Calibrated and mosaiced images, extracted spectra, catalogs, etc.
- Staged in the cloud and co-located with significant computational resources
- Open source, modular imaging pipeline facilitates custom reprocessing

Data storage & processing

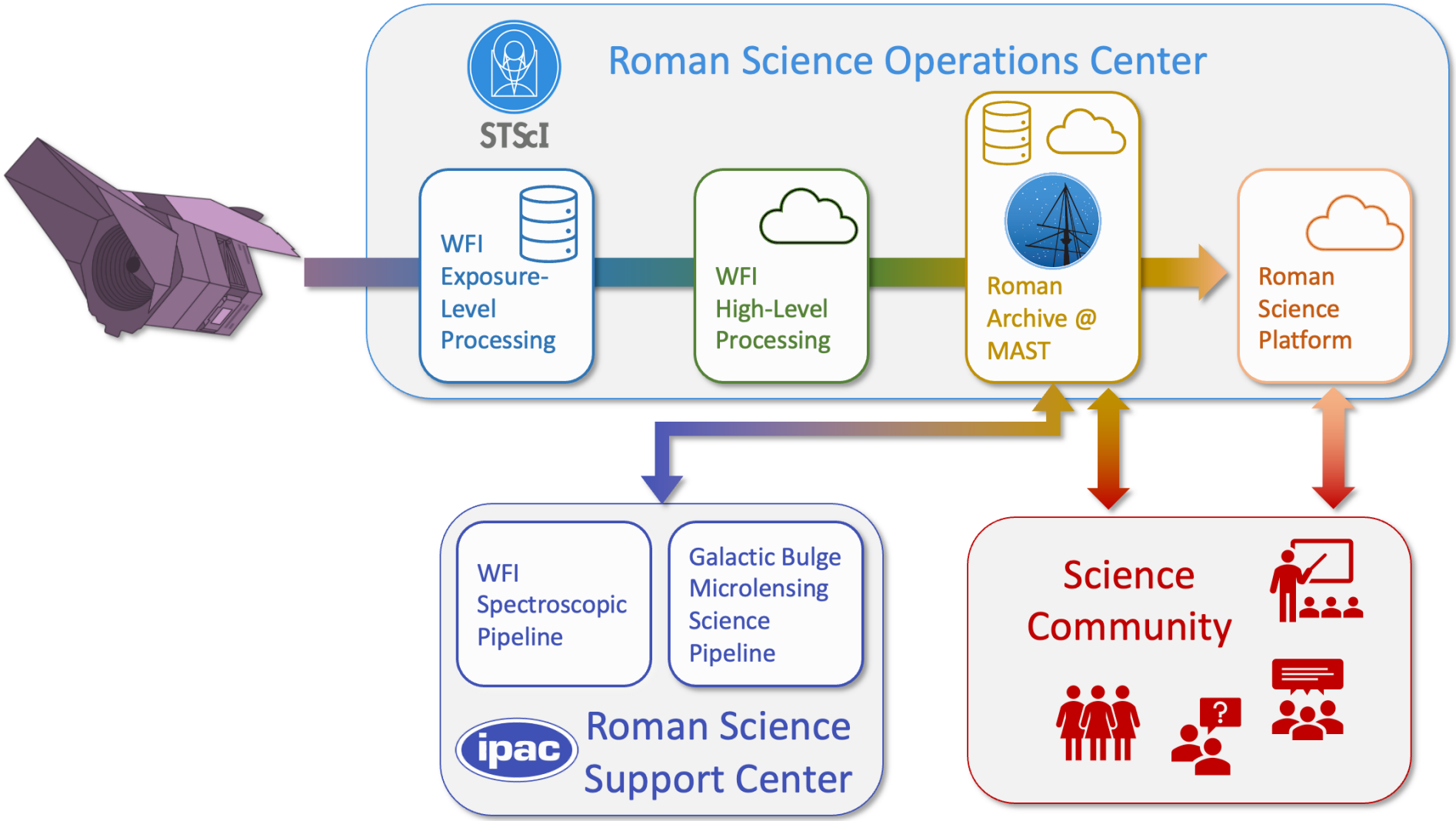
- Cloud-based high-level data processing brings software to the data
- Jupyter Lab environments ease access, sharing and repeatability
- Software environment for the community in sync with mission data processing

AWS

jupyter

STScI | SPACE TELESCOPE SCIENCE INSTITUTE

WFI Data Flow

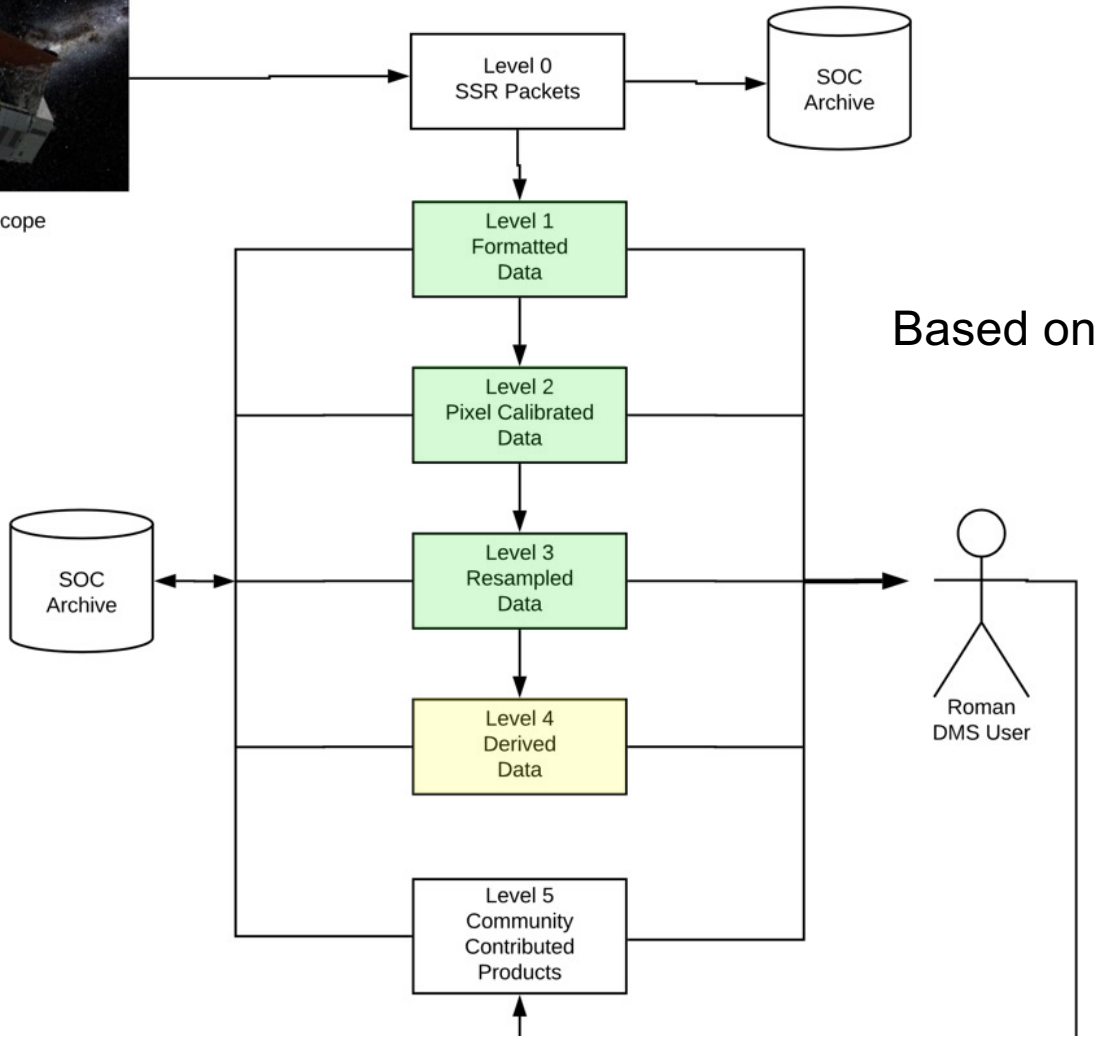


SOC Pipeline Levels

Roman Space Telescope

PIT/Community Involvement

Based on JWST



Level 1 to Level 2 Data Processing

Step Name	Roman Implementation / Changes from JWST
Data Quality Initialization	• Applies bad pixel mask to science data. • Preserves raw reference pixel values up the ramp for analysis by science teams.
Saturation Flagging	• Flags saturation based on read pattern (MA table).
IPC Correction	• Removed.
Superbias Correction	• Implementation details pending WFI FPS and TVAC data analysis.
Reference Pixel Correction	• Uses the Improved Roman Reference Correction (IRRC) algorithm from Rauscher et al. (similar to NIRSpec IRS2)
Classic Non-Linearity Correction	• Same as JWST baseline.
Dark Correction	• Same as JWST baseline. Reference files are pre-computed to match MA tables rather than re-computed on-the-fly to speed up processing and reduce reference file sizes.

Implemented | In Development | Removed | New to L2

Level 1 to Level 2 Data Processing

Step Name	Roman Implementation / Changes from JWST
Persistence Correction	• Removed
Jump Detection & Ramp Fitting	• These steps have been combined. Generalized computation of the weights handles unevenly sampled ramps (Casertano 2022).
Gain Scaling	• Removed. L2 products are stored in instrumental units of data numbers (DN) / second. Gain information planned to be stored in L2 metadata (details TBD).
Background Subtraction	• Removed. Backgrounds are not removed in L2 products to preserve noise information. Background matched in L3 products and removed during creation of L4 products.
Assign WCS Information	• Same as JWST baseline. This is the basic population of a WCS based on pointing and geometric distortion information.

Implemented | In Development | Removed | New to L2

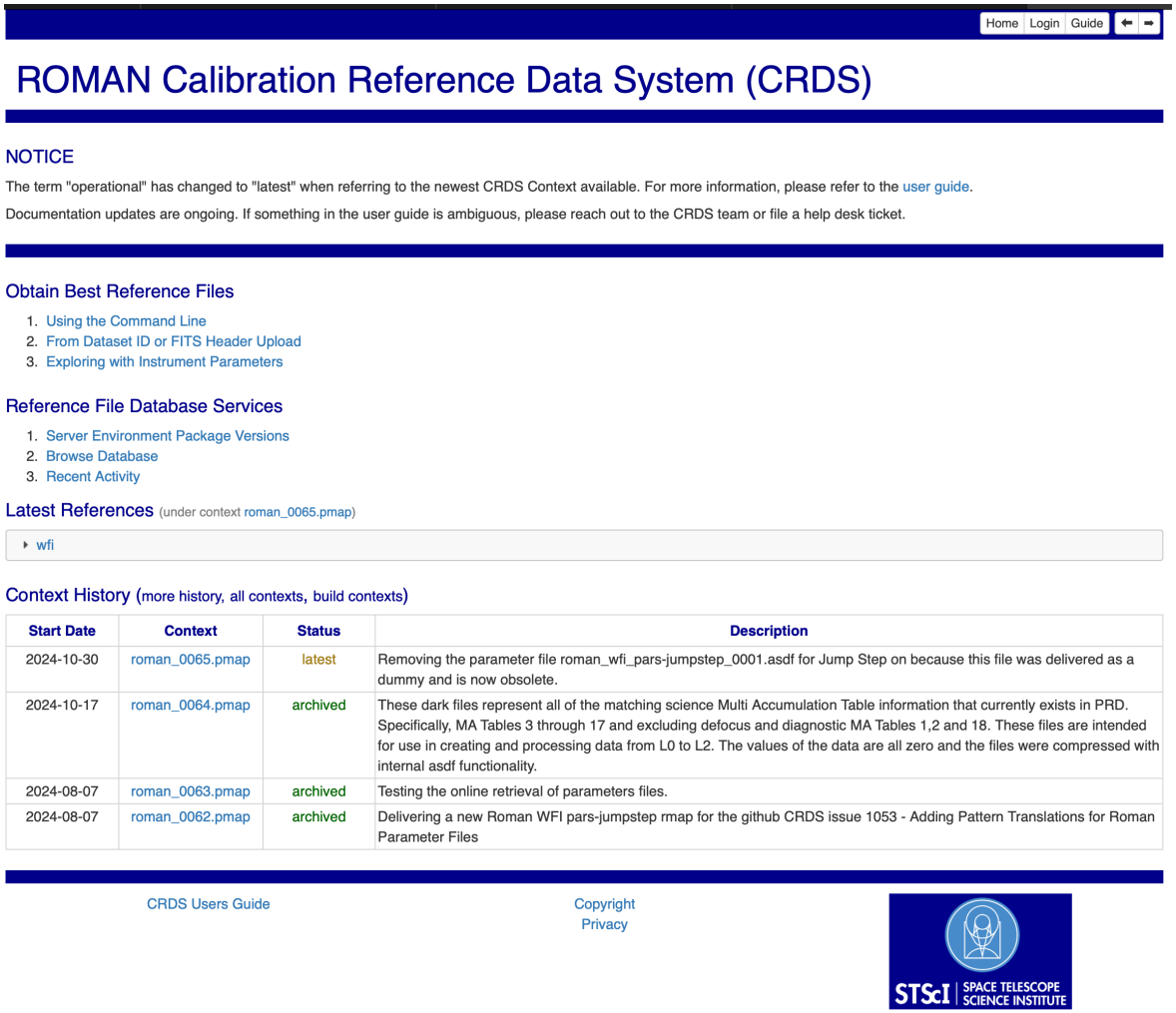
Level 1 to Level 2 Data Processing

Step Name	Roman Implementation / Changes from JWST
Flat Field	• Same as JWST baseline. • Flat fields are only applied to WFI imaging mode data. Spectroscopic data are not flat fielded at the SOC, and further calibration will be handled by the SSC.
Photometric Calibration	• Add constants to metadata that transform from instrumental units (DN / second) to physical units (both MJy / sr and µJy / arcsec^2). Unlike JWST, do not convert the pixel values.
Source Detection	• Create a source list to register the image with Gaia astrometric catalog. Use PSF fitting to determine precise positions. Some suggestions for optimization opportunities.
Tweakreg	• Compute the offsets to the WCS to match the source positions with Gaia astrometry. Wait for all 18 WFI detector files to perform a global match for the exposure. In JWST, this is performed at the start of L3 processing.

Implemented | In Development | Removed | New to L2

Calibration Reference Data System (CRDS)

https://roman-crds.stsci.edu



Home | Login | Guide

ROMAN Calibration Reference Data System (CRDS)

NOTICE

The term "operational" has changed to "latest" when referring to the newest CRDS Context available. For more information, please refer to the user guide.

Documentation updates are ongoing. If something in the user guide is ambiguous, please reach out to the CRDS team or file a help desk ticket.

Obtain Best Reference Files

1. Using the Command Line
2. From Dataset ID or FITS Header Upload
3. Exploring with Instrument Parameters

Reference File Database Services

1. Server Environment Package Versions
2. Browse Database
3. Recent Activity

Latest References (under context roman_0065.pmap)

▸ wfi

Context History (more history, all contexts, build contexts)

Start Date	Context	Status	Description
2024-10-30	roman_0065.pmap	latest	Removing the parameter file roman_wfi_pars-jumpstep_0001.asdf for Jump Step on because this file was delivered as a dummy and is now obsolete.
2024-10-17	roman_0064.pmap	archived	These dark files represent all of the matching science Multi Accumulation Table information that currently exists in PRD. Specifically, MA Tables 3 through 17 and excluding defocus and diagnostic MA Tables 1,2 and 18. These files are intended for use in creating and processing data from L0 to L2. The values of the data are all zero and the files were compressed with internal asdf functionality.
2024-08-07	roman_0063.pmap	archived	Testing the online retrieval of parameters files.
2024-08-07	roman_0062.pmap	archived	Delivering a new Roman WFI pars-jumpstep rmap for the github CRDS issue 1053 - Adding Pattern Translations for Roman Parameter Files

CRDS Users Guide

Copyright
Privacy

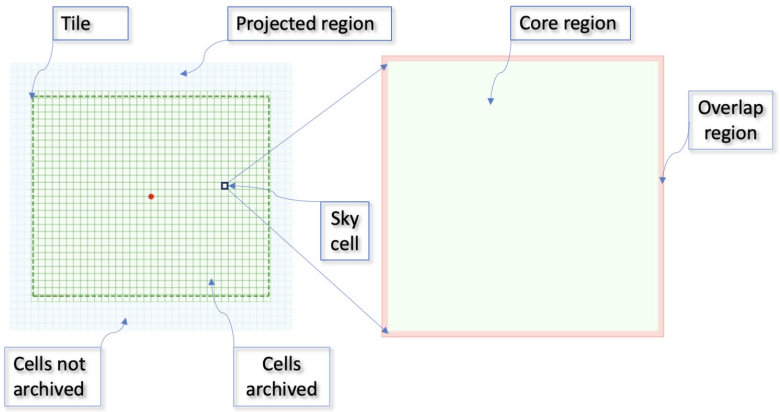
STScI | SPACE TELESCOPE SCIENCE INSTITUTE

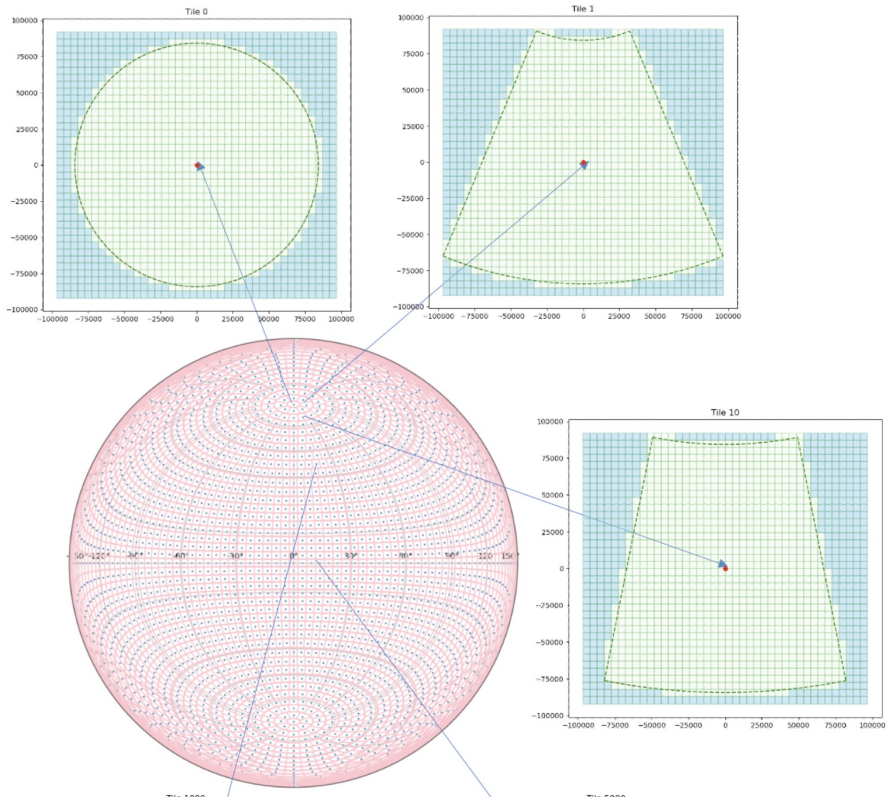
Level 3 processing

- Images are rectified (projected onto a tangent plane)
- Stacked
- Outlier-rejected
  - Unclear whether/how to save the record of the outliers
- Resampled
  - final sampling ~0.11”/pix for visit stacks, 0.05”/pix when dither pattern allows for this with no gaps
- Projected onto a grid of sky cells (based on HEALPIX)
- Several different depths of stacks are generated

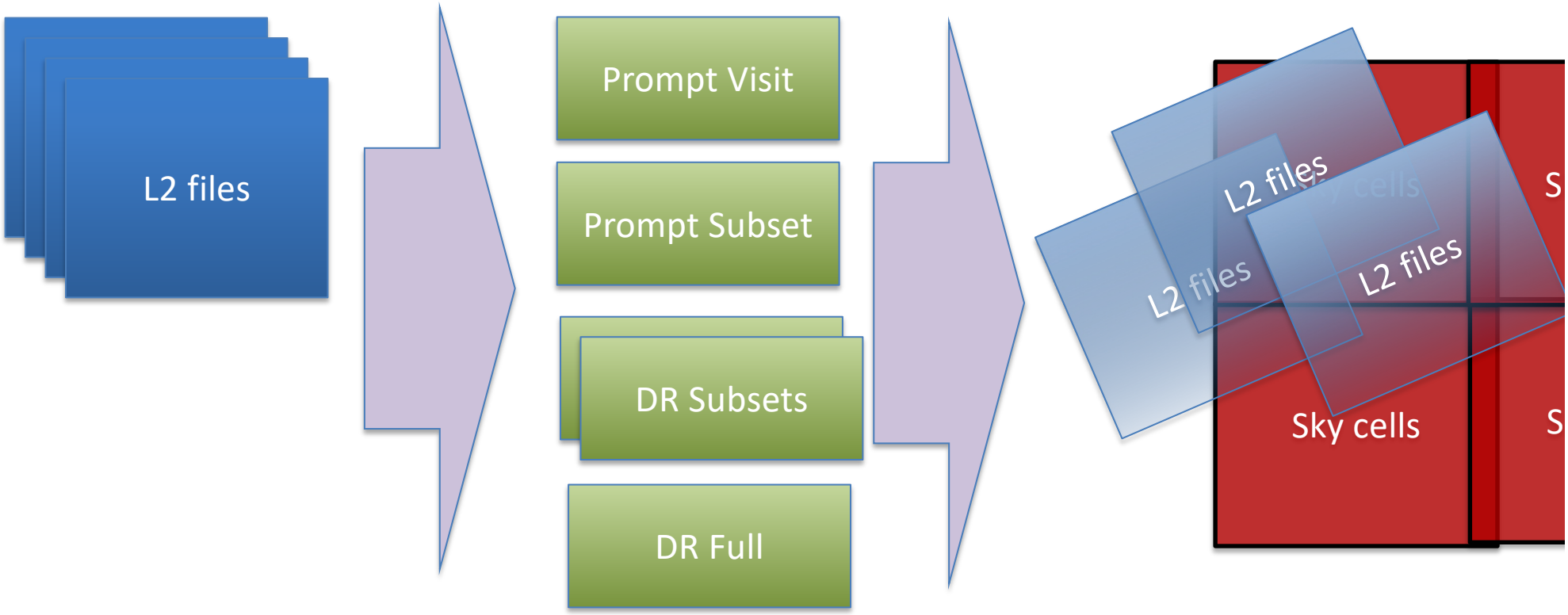
Projection Regions and Sky Cells

- Architecture is flexible & table driven
  - Supports mapping the sphere onto ***Projection Regions*** (defined by a single tangent point). The regions are then tiled with rectangular ***Sky Cells***.
  - HEALPIX-derived tiling is our baseline.
  - L3 images are on Sky Cells.





Associations



Level 3 Data Products

When	Product	Rationale
Prompt	VISIT	Useful for data-quality inspection, early science. (Native sampling. These will be patchy. Visit boundaries do not align with skycells).
Data-release	SUBSET	The subset is generally a Pass in the APT file, but this will be customized for the core surveys.
Data-release	FULL	If it differs from the Subset, this will be generated for every data release (not waiting until the end of the survey).

Level 4 processing (catalogs)

Input data	Catalog Description	Files	MAST merged Database?
L2	Individual band (primarily for image registration)	detectors	N
L3 prompt	Individual band (primarily for DQ assessment)	skycells	N
L3 data release	Multi-band (For science; photometry accurate enough for photo-z).	skycells	Y
L3 data release	Source-injection input/output	skycells	Y
L2	Forced photometry (for time-domain astronomy)	detectors	Y

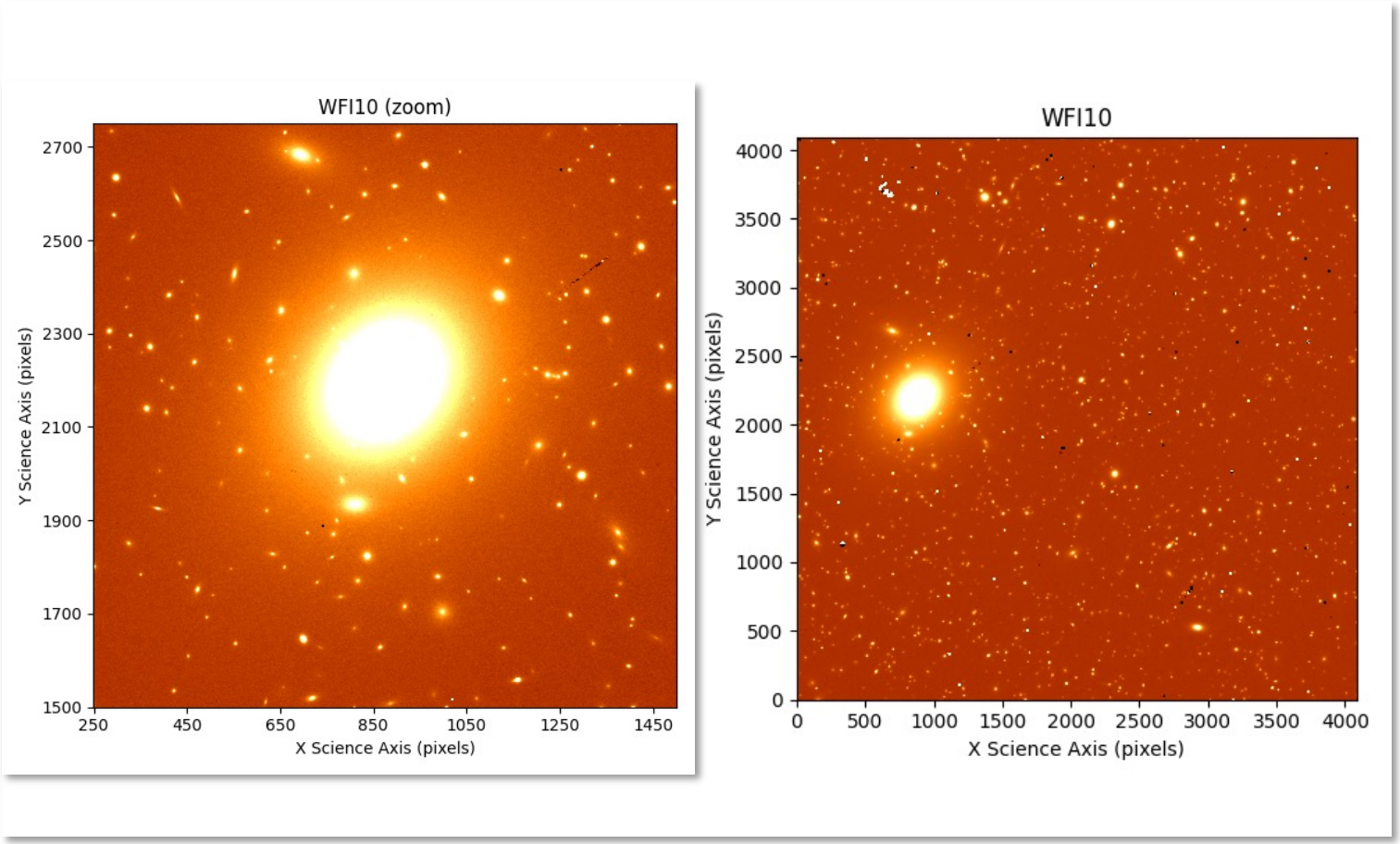
File format:

- Parquet when created
- Postgresql on premises in MAST (STScI)
- Parquet HATS on AWS S3

Working on integrating rich metadata into parquet

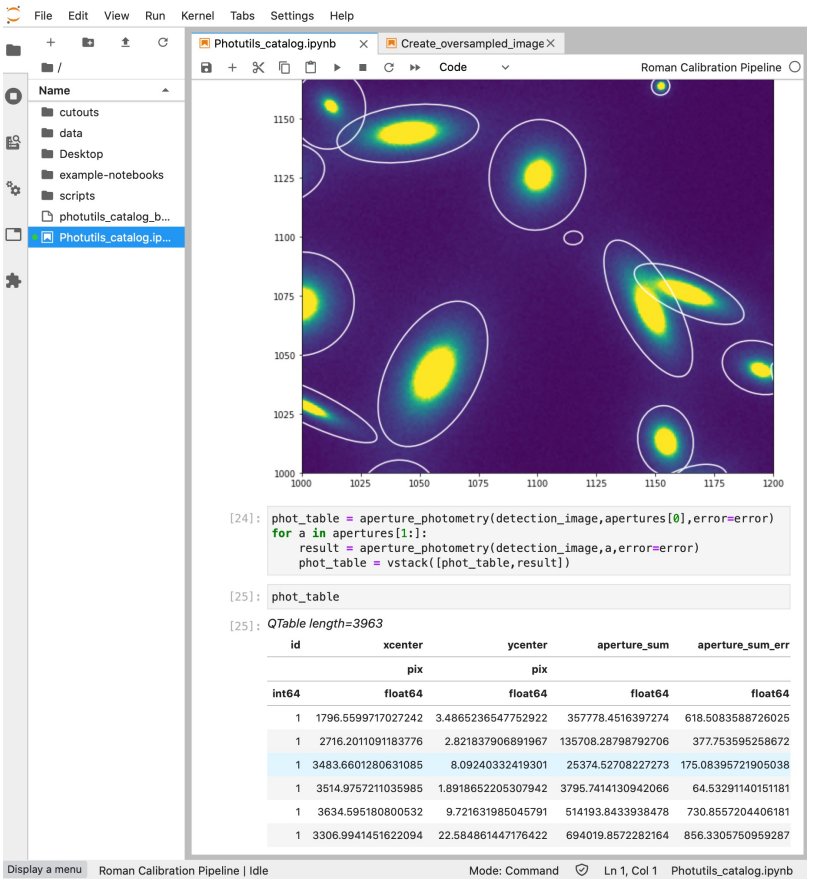
Roman I-Sim

- Roman WFI imaging mode simulator based on galsim
- Creates simulated products using SOC schema and data format compatible with romancal



Roman Research Nexus on AWS

- JupyterHub instance
  - Roman science calibration pipeline software installed and configured
  - Full Python + Astropy ecosystem installed and configured
  - Ability to install other packages and your own code
- Flexible, scalable architecture
  - Simple to add CPU & storage
  - High-throughput access to the data
  - Can scale up resources (e.g. GPUs or neural engines) as science needs & technology evolves
- No proposal required for access
  - Tiered resource allocations



Vision for the Research Nexus

The Roman Research Nexus will enable the community to perform **transformational science** on big data by providing a rich computing environment that will allow **broad, low-barrier access** to **data, compute, and software** resources.

*Enable **collaboration***

- Shared disk space and computing resources within private groups
- "Real-time collaboration" tools on shared servers

*Support the **community***

- Training: documentation, tutorials, classes
- Communication: helpdesk, email, notifications
- Enhanced user experience/interfaces

*Build **software and tools** that bring people to the RSP*

- Software for discovering, calibrating, analyzing and visualizing *Roman* data
- Cloud-native services (cutouts, catalogs)

Interoperability Issues:
Advanced Scientific Data Format

- Used by JWST and DKIST
- Hierarchical and human-readable metadata in YAML format
- Efficient binary array storage with support for memory mapping and flexible compression.
- Content validation using schemas (using JSON Schema)
- Native and transparent support for most basic Python data types, with an extension API to add support for any custom Python object.
- Explicit versioning.

Greenfield, P., Droettboom, M., & Bray, E. (2015)

ASDF: Future directions

- Support for widely used community tools (e.g. Image viewers, SExtractor, …)
  - Adding support for ASDF to Aladin Lite sky viewer
- Improved C, C++, Java support
- Improved cloud support
  - Python library supports fsspec which makes range requests on S3
  - Increasing support for zarr and dask
- Improved archiving and discoverability of schemas

World Coordinate Systems

- Open-source gwcs library used by JWST & Roman
  - Data models serialized in ASDF
- Invested some effort to try to build interoperability with LSST/Rubin
  - Did not achieve consensus on implementation of IVOA working draft (2020); development teams lacked resources for ambitious approach.
  - Developed high-level Astropy WCS abstraction layer
- Could benefit from investment…

Sky Tesselation

- LSST & Euclid were already going different directions
- Not likely that we could make a tessellation work that would avoid having to reproject (e.g. choosing the same tangent plane centers & pixel scales that were integer multiples).

Parquet files & Metadata

- Looked at IVOA Note on parquet
  - Proposes serializing VOTable XML metadata into the parquet footer.
- Decided to use YAML serialization of Astropy Table metadata (also what LSST is doing).
- Still trying to figure out full workflow for schemas & validation
  - Fuctionality that was built into ASDF not currently there for parquet

E.g. Ability to build documentation from the schema files themselves…

blackbody-1.0.0

Blackbody model.

Description

Blackbody model using the Planck function.

$$B_{nu}(T) = A\frac{2h\nu^3/c^2}{exp(h\nu/kT) - 1}$$

Outline

Schema Definitions

This node must validate against **all** of the following:

- transform-1.2.0
- This type is an object with the following properties:
  - **scale**

    object *Required*

    Scale factor.

    This node must validate against **any** of the following:

    - ../unit/quantity-1.1.0
    - **number**
  - **temperature**

    ../unit/quantity-1.1.0 *Required*

    Blackbody temperature.

Examples

$$B_{nu}(T) = 10.0\frac{2h\nu^3/c^2}{exp(h\nu/k * 6000) - 1}$$

```
!transform/blackbody-1.0.0
  scale: 10.0
  temperature: !unit/quantity-1.1.0 {unit: !unit/unit-1.0.0 K, value: 6000.0}
```

Original Schema

```
%YAML 1.1
---
$schema: "http://stsci.edu/schemas/yaml-schema/draft-01"
id: "http://stsci.edu/schemas/asdf/transform/blackbody-1.0.0"
title: >
  Blackbody model.
```

One more item on the wish list

- Point-Spread Functions
  - Universal API to get a PSF image at desired sampling?
  - Universal API to make PSF matching kernels?

Resources

Romancal Repository
github.com/spacetelescope/romancal

Roman I-Sim Repository
github.com/spacetelescope/romanisim

Roman User Documentation (RDox)
https://roman-docs.stsci.edu

Roman SOC Help Desk
https://romanhelp.stsci.edu