DSA-2000 Document No. 00023

RCP to DAT ICD

Casey Law, Martin Pokorny

Caltech

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Version: | 1 | | Version date: | 8/28/23 | | Original date: | 2023-08-28 | | Controlled document: | No | | WBS Level 2: | DAT–Data Management | | Document type: | ICD–ICD  ICD | |
|  |

Revision History

| **Ver.** | **Date** | **Sections Affected** | **Reasons / Remarks** | **Author(s)** |
| --- | --- | --- | --- | --- |
| 1 | 2023-09-01 | All | Original | Casey Law |

Table of Contents

[1 Introduction 4](#_Toc438804585)

[2 Logical Interface 4](#_Toc74243832)

[3 Physical Interface Options 4](#_Toc1947882240)

[3.1 In RCP Memory 4](#_Toc1172033440)

[3.2 At Storage System 4](#_Toc944547563)

[3.3 How to Choose 5](#_Toc1599626878)

[4 Data Formats 5](#_Toc1967666570)

[4.1 Images 5](#_Toc1966235928)

[4.2 Pulsar Timing 5](#_Toc1816636270)

[4.3 Fast Time Domain 5](#_Toc1774953333)

[4.4 Pulsar and FRB Search Metadata 6](#_Toc1611800690)

[5 Handshake 6](#_Toc1741020748)

Abstract

The RCP subsystem generates raw data that the DAT subsystem receives and processes into science-ready products. This interface must meet requirements on a sustained rate of recording data, as well as having data formats defined such that RCP can write and DAT can read at a high rate. This document defines that interface, including the logical separation of software development concerns between “radio camera” and “post-processing” pipelines.

# Introduction

The real-time processing of DSA-2000 data includes both a low-latency (sub-second correlation timescale) RCP subsystem and high-latency (>10 min pointing timescale) post-processing system (PPS; within DAT). The RCP generates data for DAT with parallelism over frequency and data type. For each stream, PPC uses custom software pipelines to perform quality assurance, mosaicking, deconvolution, and other kinds of processing. This includes processing of images, as well as outputs from Fast Time Domain (FTD) RCP pipelines for pulsar and FRB science applications.

# Logical Interface

There are two options for the implementation of the interface between RCP and DAT. The choice of implementation will depend on whether the data quality produced by the RCP. The design choice will be studied prior to FDR and is listed in the project risk register.

Here, we take care to define a logical interface, which is at a higher conceptual level than the implementation. This logical interface allows us to separate concerns between RCP and DAT.

Logically, we define the **end of the RCP** as the output of the RCP pipelines. These pipelines have a strict requirement for low latency processing, where “low” is defined by the time required to process the data stream with the memory available on a single GPU. The pipelines differ by data product (see Section 4), but are broken into two broad categories with distinct output after their final processing step:

1. Field images – an image of the sky for a specific frequency range, number of channels, and number of polarizations. Effectively, the last step is to apply an FFT to convert gridded visibilities into a sky-domain image.
2. FTD output – one or more calibrated, tied-array beam. The last step is to provide data to DAT to reproduce the signal identified by the low-latency RCP pipeline.

Logically, the **start of DAT** is defined as the input to the DAT pipelines. These pipelines have looser latency requirements, more complex processing, and higher quality requirements.

In both cases, the RCP and DAT pipelines send metadata to an observing database that describes the state of data as it is processed (e.g., has a field been deconvolved, does it pass field quality checks, etc).

# Physical Interface Options

## In RCP Memory

The baseline design will implement the interface in fast read/write memory (e.g., TB-scale SSD drives) on the RCP nodes. Moving the RCP/DAT interface to within the RCP node memory space allows higher I/O transfer. The DAT post-processing reduces the data through spatial averaging (mosaicking), which reduces the data rate output to the on-site storage system.

In this scenario, the RCP nodes will host both RCP and DAT software. The end of RCP processing will trigger the execution of DAT processing that is required to form mosaics.

## At Storage System

An alternative RCP/DAT interface may be defined to be at the on-site storage system. The storage system is accessible to the RCP software that writes data and the PPS that reads data.

In this scenario, the data will be written in parallel fashion, as each radio camera node processes a range of channels of a specific radio camera pipeline (e.g., continuum vs zoom1 vs zoom2). Each image (per channel range per pipeline type) will have a dedicated process to write data to disk, which defines the end of the RCP.

## How to Choose

There are hardware cost savings to having the interface within an RCP node, but this comes at the cost of software complexity and data quality risk. A design study will use forward modeled field images to define the quality of mosaics under varying assumptions and kinds of processing. This will define the role of the DAT processing that must occur on the RCP node and whether it hinders RCP processing.

# Data Formats

All data format details are subject to change as software is developed and tested. We include notional data formats initially.

## Images

This format is defined to hold continuum, polarimetric, and all spectroscopic RC image outputs. The memory limits on the GPU (nominally ~20 GB) require that a small number of channel-polarizations (nominally 4x4) are processed by each GPU. On a single RCP node, more channels will be grouped and mosaicked in some fast, local memory.

The standard survey mode will generate RCP image output every 10.3 minutes. For reference, the size of a 16k square channel-pol image is 1 GB (single-precision, 32 bits per pixel). The RCP output channel counts and resolutions are:

* 5492, 130-kHz, Stokes I channels, fixed selection
* 625, 2.0-MHz, Full Stokes channels, full band
* 4192 8-kHz, Stokes I channels, fixed selection
* 960 1-kHz, Stokes I channels, tunable selection

Out of this, the DAT subsystem will create the full suite of mosaicked, science-ready images, including averaging over channels.

The RCP will provide a PSF to the post-processing system or information required to generate a PSF. The correctness (I.e., derived from data or pre-calculated) is not yet defined. The grid size will be fixed, so the field of view, PSF, and pixel size will scale with frequency (e.g. pixel size 1.93” at 0.7 GHz, 1.0” at 1.35 GHz, and 0.675” at 2 GHz).

## Point Spread Function

The RCP will know how visibilities are gridded and thus is the only subsystem with knowledge required to generate the Point Spread Function (PSF) of an image. The PSF will be provided by RCP as an image or a function to generate such an image. The PSF must be large and accurate enough to reveal sidelobes that are 5e5 times smaller than the brightest image in a typical field (I.e., 2 microJy and 1 Jy, respectively).

## Pulsar Timing

This format is defined for millisecond pulsar timing, which will record data from up to four beams at any given time.

In millisecond pulsar timing, the data is referred to as “Stokes profiles” and defined to hold tied-array beam power data folded at an assumed dispersion measure and period. A standard Stokes profile will be saved each 10 s with 2048 phase bins across pulsar rotation period and repeated during typical target session.

The baseline plan is to use the PSRFITS format to hold pulsar timing data. Header and data shape details will be defined here later.

## Fast Time Domain

This format is defined for FRB search and pulsar search. Both types of data are not recorded continuously but based on real-time decisions made by the search systems in RCP pipelines. This is referred to as “triggered” data recording.

For pulsar search, the RCP output data is a power beam on target with 0.2-ms resolution (~10 min duration) and retaining frequency and polarization axes. Provisionally, we will save this in PSRFITS format. [An alternative is to save folded data with axes of time, phase, channel, polarization]

For FRB search, the RCP output data is a voltage beam on target [what duration?] that has been dedispersed and retaining polarization axes. Provisionally, we will save this in PSRFITS format. In addition, a custom data product (format TBD) will save a dedispersed Stokes I image cube toward the target that retains time and frequency axes. This will be used for precision localization of the FRB.

The pulsar timing and FTD data sets have one common need, which is to save a standard tied-array, time-domain product in PSRFITS. All three of these products have different assumptions (folded vs non-folded, axis shapes), but the goal is to use a common data format when possible.

## Pulsar and FRB Search Metadata

Each pulsar timing and FTD data set will be associated with metadata and a visualization to aid human interaction and quality assurance.

For pulsar search, candidates are identified in beam-formed power dynamic spectra via a folding algorithm. Nominal parameters of the search are DM, period, and period derivative. Additional information includes: SNR. This list will be expanded as the software is developed.

For single-pulse search, candidates are identified in beam-formed power dynamic spectra. Nominal parameters of the search are DM, time, temporal width. Additional information to aid human interaction and quality checks include: SNR, beam location.

These metadata can be saved in json format locally and loaded to the metadata database. The database will keep information about data products associated with trigger information.

# Handshake

The interface between RCP and DAT will require a “handshake” to acknowledge that the transfer was successful and that DAT is responsible for the data. The metadata database exists to track the state of data as it comes out of the RCP, through DAT, and into ARC. This makes the database a natural place to register a handshake after the RCP completes its processing.

For images, the handshake will reference a unique identifier for a field data set, which could include information such as time and position on the sky. For FTD data, each trigger is associated with a packet of data and metadata. The trigger will have a unique identifier that will be associated with the data and metadata via the metadata database managed by DAT.

In addition, any RCP output will have one or more summary statistics calculated from the data and pushed to the database. The DAT system will be able to verify the data by recalculating the statistics.