DSA-2000 Document No. 00022

Post-Processing to Public Archive ICD

Casey Law, Rachel Akeson

Caltech, IPAC

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | 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 and good draft | Casey Law |

Table of Contents

[1 Introduction 4](#_Toc1228111045)

[2 Interfaces 4](#_Toc1066630451)

[2.1 Suitcase 4](#_Toc1124251036)

[2.2 Remote access 4](#_Toc1963551887)

[2.3 Quality Assurance 4](#_Toc1913621933)

[2.4 Handshake 5](#_Toc694444222)

[3 Contents of Suitcase 5](#_Toc1966199627)

[4 Links to data fomats 5](#_Toc828567942)

Abstract

We define the concept for moving data from the on-site post-processing system and the public archive. The interface centers on the idea of a “suitcase” of data, which may be transferred via fiber or shipped disks.

# Introduction

The post-processing system applies certain corrections and quality control to the raw output of the RCP. The data processing building located on site will run RCP and post-processing computing systems, as well as hold data during post processing (referred to as ‘on-site’). Off-site storage includes a redundant copy of the on-site data and data sent to the public archive for further processing and serving to the public. This document defines the interface between the post-processing system and the public archive.

Internet access to the site will define the maximum data throughput and transfer costs for this interface. The operational concept is also coupled with the internet access, as some access options require physical labor. The baseline requirement is built around an assumed fiber backhaul connection that can support transfer of 100 Mb/s averaged over a full year (up to 3.1 PB/yr).

# Interfaces

## Suitcase

The “suitcase” is a quantum of data from a fixed amount of observing time and fixed post-processing and Quality Assurance (QA). A suitcase of data will have uniform processing by the public archive software and its quality control systems, then be released to the public. The baseline requirement is for three public releases. While the releases are batched for quality control purposes, the suitcase may be delivered either as a batch (i.e., via shipment of disks) or steadily (i.e., via the internet).

The steady transfer of data can be managed by rsync or a similar tool. This would allow control over transfer rate, as well as features to retry failed transfers.

For batch transfer via disks, staff will manage copying data to portal disk packs on site. Disk management is difficult and risks damaging hardware and losing data. Therefore, the cadence of shipping disks would be limited to less often than once per month and more often than the release cadence baseline.

For either transfer mechanism, the process will be tracked via the metadata database, which manages the state, quality, and location of data output by the RCP.

## Remote access

A second interface provides users off site with the ability to “pull” data to them directly from the on-site archive. This is mediated by the public archive and uses tools built by IPAC to access data installed to access on-site data. Currently, the integration of the public archive with on-site data, including remote data access, is a “desirement”.

## Quality Assurance

In order for data to qualify for transfer through this interface, it must pass quality assurance (QA) tests in the post-processing system. The metadata database will manage the state of data quality, as assessed by the post-processing subsystem.

The post-processing QA will be designed such that:

1. Files are in standard format with metadata required for their interpretation.
2. Transferred data have scientific value, as defined by the science requirements.
3. Public archive software can process data.

The public archive subsystem will also assess quality in its own pipelines. These QA results will be shared back to the metadata database.

## Handshake

A “handshake” is a procedure to verify that a transfer was completed successfully. After the handshake completes, the public archive may proceed with its processing of the data with confidence that it is valid and consistent with that sent by DAT.

The handshake will depend on the nature of the transfer. The process will be exercised and documented here when finalized.

# Contents of Suitcase

The suitcase will be delivered from the on-site archive to IPAC. The baseline definition of the suitcase is as follows.

* Wide Continuum and Full Continuum:
  + either 1 (1 microJy) or 5 to 6 (2 microJy) epochs of continuum Stokes I images
  + 0.56 Terapixel/sky at 0.81” resolution (2.25 TB/sky)
  + 11 image planes (10-quasi-continuum + full-band)
  + 25 TB per epoch
* Narrow Line:
  + 6e5 million galaxies
  + 9-point grid
  + 500 frequency channels
  + 11 GB
* Zoom A:
  + 3e6 galaxies
  + 120x120 to 1200x1200 pixel image
  + 100 frequency channels
  + 31 TB
* Zoom B:
  + Milky Way low latitudes (b<~10)
  + 960 frequency channels
  + 50 TB
* Medium Polarimetric:
  + 6e7 9-pt cutouts around 20 sigma Stokes I sources
  + 605 frequency channels in Stokes IQUV
  + 5 TB
* PTA:
  + Not included in suitcase
  + PTA data gets copied to NANOGrav archive
* Periodic timing confirmation:
  + Dedispersed, power beam with 0.1 ms resolution and folded at candidate period and DM
* FRB:
  + Dedispersed, full-Stokes, power beam with xx ms resolution at candidate DM and location.
  + Localization via cutout image cube.

# Links to data fomats

The contents of the suitcase will be structured as multiple files, according to what is easy to produce, transfer, and process by the public archive systems. The intention is to use common data format standards, such as FITS, PSRFITS, and hdf5.

The formats will be defined during development and documented here.