Title: SRDP Transition Plan	Owner: Lacy	Date: 1/31/2023
Document No.		Version: 0.1

Science Ready Data Products

Plan for transition to Operations
Draft 0.2

PREPARED BY	**ORGANIZATION**
Mark Lacy, John Tobin	NRAO
VERSION RELEASE BY	**ORGANIZATION**
Mark Lacy, Operations Manager	NRAO

APPROVALS	**ORGANIZATION**	**SIGNATURE**
Anthony Remijan	NRAO SSR AD and SRDP Project Director	

Title: SRDP Transition Plan	Owner: Lacy	Date: 1/31/2023
Document No.		Version: 0.1

Change Record

VERSION	DATE	RESPONSIBLE	REASON
0.1		M. Lacy	First daft

Title: SRDP Transition Plan	Owner: Lacy	Date: 1/31/2023
Document No.		Version: 0.1

Table of Contents

1 Introduction

2 Background

The SRDP project (in collaboration with existing observatory management structures) will develop and implement the software tools, scientific heuristics, and operations structures to deliver science quality data products to the NRAO user community, both as it exists now, and the foreseen expansion of the community to include astronomers who do not have a detailed knowledge of radio astronomy.

2.1 Scope of this Document

The SRDP project was conceived as a 5-year effort, beginning in FY2019. After the completion of the project, the capabilities developed will be turned over to the operations teams in Socorro and Charlottesville (in practice, some capabilities remain under development, so turnover of them may be delayed beyond the nominal project completion date). In this document, we describe the conditions needed to transition an SRDP capability to operations.

2.2 Reference Documents

[RD01] SRDP Project Charter 530-SRDP-001-MGMT

[RD02] SRDP Project Management Plan 530-SRDP-003-MGMT

[RD03] SRDP System Engineering Management Plan 530-SRDP-010-MGMT

[RD04] SRDP System Concept Document 530-SRDP-014-MGMT-SRDP_System_Concept

[RD05] SRDP Operations Plan 530-SRDP_Operations-plan

Lacy et al. 2020: The VLA Sky Survey

Project documents may be found on Sharepoint:
https://sharepoint.nrao.edu/pmd/projects/530%20Science%20Ready%20Data%20Products

2.3 Abbreviations and Acronyms

ALMA – the Atacama Large Mm/submm Array
API – Application Program Interface
ARC – ALMA Regional Centre
AUDI – ALMA user-driven imaging
AWS – Amazon Web Services
DMS – NRAO's Data Management and Software department
HTC – HT Condor job management software
JAO – Joint ALMA Observatory
MS – Measurement Set
NA – North America
TTA tools – next generation proposal submission/review tool
QA – Quality Assurance
SB – Scheduling Block
SRDP – Science-Ready Data Products
VLA – the Karl G. Jansky Very Large Array
VLASS – the VLA Sky Survey

VUDI – VLA user-driven imaging
WF - workflow

2.4 Assumptions

2.4.1 SRDP Use Cases/capabilities

The SRDP System Concept Document [RD04] identifies several use cases, that we have mapped into capabilities (or workflows; WFs)[1] for support:

- WF1: Standard Calibration (VLA only): running the calibration pipeline until satisfactory, calibration tables are produced, then archiving.
- WF2: Standard Imaging (VLA only): restoring calibration tables, using cached calibrated measurement sets on disk or in a single recipe with WF1, flagging target data as needed, then running the imaging pipeline until satisfactory images are produced, then archiving.
- WF3: Optimized imaging (VLA and ALMA): restoring calibration tables, running the imaging pipeline with user-specified parameters, archiving products that pass QA.
- WF4: Archive (VLA and ALMA): data discovery and download, visualization.
- WF5: Restoration of Calibration (VLA and ALMA): including small modifications such as smoothing in time or frequency, generation of MS etc)
- WF6: Recalibration (VLA and ALMA): running the calibration pipeline with user-specified parameters (including line emission information so as to avoid flagging spectral lines as RFI, until these can be specified in the ngPST), then archiving products that pass QA.
- WF7: Combined imaging (VLA and ALMA): restoring calibration tables, running the imaging combination pipeline with user-specified parameters, archiving products that pass QA.
- WF8: Time Critical observations (VLA only): for pre-identified SBs, running the calibration and imaging pipelines (for a restricted set of images) on individual execution blocks as soon as they are observed, without waiting for the full set of SB executions to be completed, and making the results available to the PI.
- WF9: Large & Commensal projects (VLA and ALMA): ingest of user-supplied products into the archive following QA, eventually special processing
- WF10: Curation and reproducibility (relevant for all the above workflows, but not a workflow in itself). (These topics will be the responsibility of the SRDP project, and will be achieved via curation of the appropriate metadata in the archive.)
- WF11: Commissioning and validation of SRDP workflows and pipelines (VLA and ALMA).

The original concept included both continuum and spectral line datasets in the VLA workflows. We have split these (WF1a, 2a for continuum, 1b, 2b for line) as spectral line capabilities are significantly more complicated than continuum and will be delivered later.

[1] Originally SRDP capabilities were referred to as workflows, however, workflows has a different meaning for the software engineers, so we use capabilities except where needed for consistency with earlier documents.

3 Capability acceptance requirements

3.1 Workflow process monitoring

The status of capability requests is given in the workspaces interface, but this is typically not fine-grained enough to diagnose problems with the workflows. The WS 2.8.1 release allows more granularity of the view on workflow progress, and this information will appear in the WS interface in WS 2.8.2. In particular, it will be possible to see when a job has completed data fetching from NGAS, which currently has a thread limit preventing more than ~10 jobs fetching data at once. Specific HTC workflows can be checked with the workflow inspector (dsoc-prod/workspaces/sbin/wf_inspector), which allows login to the running job by providing a wrapper to the HTC ssh_to_job command. Ultimately, a queue manager will automatically manage job submissions and prevent overload of the NGAS thread count.

3.2 Reliability

Minimum reliability requirements on the successful execution of a workflow (from initiation to archiving of the final products) will be set, which need to be met as a precondition before a workflow is handed off to operations. We propose 99% reliability for routine pipeline capabilities (not counting failures due to observational setups that are incompatible with the pipelines) and 90% reliability for user-driven workflows.

3.3 Analyst workload

Before handing a workflow over to operations, an assessment will be made of the ongoing need for data analyst and scientist effort. For example, 2.3 pipeline runs (an initial run and 1.3 rereuns) are needed on average to deliver a VLA SRDP calibration product, which is a significant load on the data analysts. Most reruns are for gain compression, RFI or DTS issues. With improved pipeline heuristics we expect this to reduce. Improved automation of calibration and imaging QA will also reduce the load on the analysts. The level of improved heuristics and automation needed for a given capability will vary depending on the analyst resources available to the respective pipeline groups and operations teams. Negotiations with the operations team are therefore needed before a capability is handed over.

3.4 Documentation

Workflows should be documented (including edge cases). QA standards should be well-established and documented in per-capability memos (in collaboration with the operations teams). For calibration workflows, the QA memos should specify the expected flux calibration accuracy, phase stability and amplitude gain stability (taking into account, for example, gain compression, decorrelation and antenna pointing).

3.5 Metrics

Workflow metrics should be easily obtainable via a documented procedure, and contain enough granularity to allow reports to senior management to be constructed. This is currently implemented via the ws_metrics task, however, additions and improvements to the reporting will be needed before we can move from the current system based on Google spreadsheets. In the long term, a graphical interface to the workspaces database is preferred.

4 Status by capability

4.1 WF1 and WF2 (VLA std cal and imaging)

Missing:

1. Ability to ingest the weblog from a failed run (descoped from WS-214; WS-1616).
2. Ability to search internal notes for recurring issues (WS-1269)
3. Spectral line calibration (avoidance of flagging lines as RFI), spectral line imaging (may be hard without prior knowledge of line frequency, unless automated line identification developed or narrow spws simply assumed to be line spws). Needs pipeline tickets, possibly also a new workspaces capability. Full implementation needs both CASA and pipeline development. CASA implementing the uvgrid flagging needed to support RFI rejection, CASR-496 and related tickets.
4. Implementation of Variable flux calibrator models for high frequency observations with 3C48 & 3C138.
5. Ability to handle limited archive threads (queue manager & NGAS upgrade).
6. Self-calibration in VLA imaging (scheduled for the Fall 2024 pipeline release).
7. Running calibration and imaging as a single capability (scheduled for WS2.9).

Continuum calibration and imaging could be moved to operations if 1-2 are implemented and (5) handled via more granularity in job reporting accompanied by manual queue management, though warnings need to be issued for flux scales until (4) is fixed and information added to qa_notes. Finally, (7) is needed to improve efficiency and provide better feedback to the analysts on the need for flagging. Also, a clean-up of the WS interface that does not show redundant or null workflows in the default menu would be useful.

4.2 WF3 (AUDI & VUDI)

VUDI in development, AUDI missing:

1. Ability to ingest more than one image per MOUS/Source combination without manual workarounds.
2. Ability to choose the pipeline version for calibration restore version automatically (needs original pipeline version to be stored in the archive).
3. Fix to session mapping bugs (to be implemented in Workspaces)

4.3 WF4 (archive downloads) and WF5 (restores)

Already in operations.

4.4 WF6 (recalibration), WF7 (multi-array imaging)

Not yet in development

4.5 WF8 (time critical VLA processing)

Missing:

1. Automated identification of time-critical EBs (needs TTA tools).

Could be put into operation with regular continuum cal and imaging using prior knowledge of time critical observations, but as yet there is no formal way to transmit this information from the user to the data processing team. There would also be a 1 working day delay as accepting and archiving is currently a manual process, though automating the ingestion process based on a "time critical" parameter in the

capability metadata would be possible. Another need would be the ability to ingest an improved product at a later date (requires support for versioning in the archive).

4.6 WF9 (archiving of large projects "collections")

Not yet in development (VLASS catalog storage and access part of CIRADA in-kind contribution).

Capability	Status as of Q3 FY23	Earliest possible handover
WF1 (VLA pipeline calibration)	Continuum almost ready, spectral line to be developed	Q1 FY24 (continuum only)
WF2 (VLA pipeline imaging)	Continuum almost ready, spectral line to be developed	Q1 FY24 (continuum only)
WF3 (Optimized imaging, AUDI & VUDI)	AUDI needs archive improvements and workspaces conversion, VUDI in development.	Q1 FY25 (AUDI), VUDI unclear.
WF4 (Archive services)	Already in operations	N/A
WF5 (VLA and ALMA calibration restores)	Already in operations	N/A
WF6 (recalibration)	Not yet in development	Unknown
WF7 (multi-array imaging)	Not yet in development	Unknown
WF8 (time critical VLA processing)	Could be manually implemented, but needs TTATools for automation.	Unknown
WF9 (large projects/collections)	In development	Unknown

Table 1: summary of capability readiness

5 Preparing for ngVLA

ngVLA will provide quality-assessed, pipelined data products for standard observation modes. The infrastructure can thus be very similar to that of SRDP, though an extra level of complexity may be introduced by the need to split jobs into many distributed, parallel processes, closer to the VLASS model. The computational expense of ngVLA processing will be an important aspect of the support model, for example, it may be cheaper to re-execute an observation than perform multiple pipeline runs to deal with data issues from an initially poor observation. Criteria based on pipeline results similar to the assisted QA being developed for SRDP could be used to automatically trigger reobservations.

6 SRDP user survey

An SRDP user survey is planned for the Fall of 2023 to investigate the uses that SRDP data products are being put to. This will be sent to recent users/PIs and other community representatives. Questions will include inputs to prioritizations of any future SRDP developments.