Science-Ready Data Products

System Concept

Project 530

PREPARED BY	ORGANIZATION
Jeffrey S. Kern	NRAO, SRDP Project Director
SRDP Requirements Committee	NRAO
REVISION RELEASE BY	**ORGANIZATION**

APPROVALS	ORGANIZATION	SIGNATURE
Lewis Ball	Assistant Director Science Support and Research NRAO	
Brian Glendenning	Assistant Director Data Management and Software NRAO	
Jeffrey Kern	Project Director Science Ready Data Products NRAO	
Robert Treacy	Project Manager Science Ready Data Products NRAO	

Change Record

VERSION	DATE	AUTHOR	NOTES/CHANGES
0.01	01/17/18		Initial Draft.
0.1	02/21/18	Kern	Draft for Review
1.0	3/15/18	Kern	Incorporated feedback from NRAO Staff Review
1.1	6/29/2018	Treacy, Kern	Incorporated feedback from StRR, CoDR,

Table of Contents

1. Introduction 4

1.1 Background 4

1.2 Scope of this Document 4

1.3 Document Authors 4

1.4 Document Outline 4

1.5 Applicable Documents 5

1.6 Reference Documents 5

1.7 Acronyms 5

2. System Overview 5

2.1 Scope 6

2.2 Actors 8

2.3 External Interfaces 9

2.4 Policy Implications 10

2.5 Computing Resource Management 11

2.6 Quality Assurance 13

2.7 Summary of Use Cases 13

3. Use Cases....................14

3.1 Use Case: *Standard Calibration* 14

3.2 Use Case: *Standard Imaging* 17

3.3 Use Case: *Optimized Imaging* 19

3.4 Use Case: *Archive Use* 22

3.5 Use Case: *Restoration* 24

3.6 Use Case: *Recalibration* 26

3.7 Use Case: *Combined Imaging* 29

3.8 Use Case: Time Critical Observations 32

3.9 Use Case: *Large Projects* 34

3.10 Use Case: Curation and Reproducibility 35

3.11 Use Case: Commissioning and Validation 36

4. Overall System Description 37

4.1 Products 37

4.1 SRDP System Overview 38

4.2 The SRDP Capability Roadmap 40

1. Introduction

1.1 Background

The Science Ready Data Products (SRDP) project is an initiative designed to remove a significant fraction of overhead associated with data calibration and imaging from the user, allowing users of NRAO facilities to focus less on data reduction and more on the cutting-edge science made possible by the VLA and ALMA. As such, SRDPs are an important means of expanding the NRAO user base. Development of SRDPs is a key deliverable under AUI's Cooperative Agreement with the NSF and the Program Operating Plan and is supported by internal funding.

Science Ready Data Products is a functional definition, and part of the SRDP project is to refine the technical definition. Science Ready Data Products span the range from calibrated visibilities, through imaged and de-convolved data cubes, to "value added" products such as source catalogs and moment maps.

A number of operational concepts will be developed and described in this document. Concepts will be translated to capabilities and requirements that will be delivered over successive release cycles. A roadmap will be maintained which shall outline a progressive plan for delivering SRDP capability, with requirements defined in detail that will be included in each particular release cycle. Concept development and requirements decomposition are interdependent processes. Concepts will be iterated as requirements develop to reflect their interdependencies. A clear definition of operational concepts will also provide guidance in the development of operations planning, as well as a common understanding between stakeholders.

1.2 Scope of this Document

This document provides the overall concept for the Science-Ready Data Products project. Actors, interfaces, and relevant policy choices are listed to provide context. A set of use cases is developed to define the scope of the project and an integrated system concept is presented. A high level SRDP Capability Roadmap is also provided to denote the sequenced rollout of capability. This document does not define the system architecture or the implementation plan.

1.3 Document Authors

This document has been prepared by the SRDP Requirements Committee in partial fulfillment of the committee charge described in the Committee Terms of Reference [AD4]. The members of the committee are: Claire Chandler, Rafael Hiriart, Amanda Kepley, Daniel Lyons, Josh Marvil, Eric Murphy, and Catarina Ubach. Walter Brisken and James Robnett contributed to this document through their roles as observers on the requirements committee. Jeff Kern was acting chair during the drafting of this document. Bob Treacy (Project Manager) and Joe Parker provided system engineering input to this document.

1.4 Document Outline

Section 2 sets the external context of the SRDP project, providing the scope and objectives of the project. The actors within the system, interfaces they use, and the operational context are described. In section 3 use cases encompassing the scope are elaborated. These use cases are the primary source for stakeholder requirements for the SRDP project. In section 4 relationships between the use cases are clarified and an overall system concept defined.

1.5 Applicable Documents

AD1 – SRDP Stakeholder Register 530-SRDP-005-MGMT

AD2 – SRDP Stakeholder Requirements 530-SRDP-015-MGMT

AD3 – SRDP Lifecycle Phases and Concepts

AD4 – SRDP Requirements Committee Terms of Reference 530-SRDP-012-MGMT

[Other Applicable Documents]

1.6 Reference Documents

RD1 – SRDP System Requirements 530-SRDP-016-MGMT

RD2 – IEEE-1362-1998 IEEE Guide for Information Technology-System Definition-Concept of Operations (ConOps) Document

RD3 – IEEE-29148-2011 Systems and software engineering - Life cycle processes - Requirements Engineering (Supersedes RD2)

RD4 – SRDP System Engineering Management Plan 530-SRDP-010-MGMT

RD5 – ALMA Operations Plan (v. D) ALMA-00.00.00.00-002-D-PLA.A

1.7 Acronyms

DMS	Data Management and Software
AAT	Archive Access Tool
PPI	Post Processing Interface
PST	Proposal Submission Tool
EDP	Enhanced Data Products
SRDP	Science Ready Data Products
NMASC	New Mexico Array Science Center
NAASC	North American ALMA Science Center

2. System Overview

The SRDP project seeks to maximize the scientific impact of the interferometers operated by the National Radio Astronomy Observatory by:

- allowing current users to focus more on science and less on data reduction,
- broadening our user community by decreasing the barriers to using these instruments, and
- creating a rich archive of science ready images and products for archival study.

The use cases detailed in section 3 of this document elaborate on how users will interact and benefit from these developments. The SRDP project will define and deploy a set of tools to support users and operations staff in fulfilling these use cases. The system is defined by the sum of the list of products to be produced, the tools required to produce the products, and the processes (both automated and manual) to effectively use those tools.

2.1 Scope

The use cases described here are the primary source for the Level-0 project requirements. The SRDP project will use a rolling wave approach to requirements management and delivery (see discussion in RD4), to iteratively refine and address the requirements. This process provides the opportunity for feedback from the community throughout the project. The rolling wave approach precludes a detailed implementation plan for the entire project at this stage, but a rough roadmap assuming a five-year project is described in section 4.2. After the SRDP project concludes the generation of science-ready products will be a normal part of observatory operations, and continual improvement and development will be part of NRAO's standard operating procedures.

The SRDP project is currently focused on the development of science-ready products for the VLA and ALMA. The National Science Foundation has requested a proposal for the re-integration of the Long Baseline Observatory (LBO) into the NRAO. If accepted, re-integration would begin with the 2019 Fiscal Year (Oct 1, 2019) and SRDP development will be extended to include the VLBA at that time.

Although both ALMA and the VLA are interferometers, significant differences exist between these instruments, both in the current level of development and in the broader context of the observatory. ALMA has made significant progress in developing processes and capabilities for the generation of science-ready products, although science-ready imaging is not yet a standard. NRAO, however, is not able to unilaterally modify the tools or policies of the ALMA telescope because of the international agreements in the governance of ALMA. Calibration and standard imaging are the responsibility of the ALMA observatory, which has developed workflows and requirements independently of the SRDP project, and will continue to do so.

The SRDP project will add value for North American ALMA users by providing capabilities not delivered by the ALMA project and will not reproduce products already delivered by ALMA (e.g. science quality calibration). The SRDP project will support ALMA by extending reasonable efforts to allow reuse of capabilities by other executives, and to incorporate and re-use developments made by the ALMA project.

Like ALMA, the VLA has developed science-ready data product capabilities. The VLA pipeline already provides reference quality calibrations, and the VLA Sky Survey (VLASS) has produced over seven thousand square degrees of (quick-look) science quality images. The VLASS and SRDP projects have significant overlap, however the schedules are set by different constraints. VLASS is currently in a period of early operations, during which requirements and operational models are still being refined. During this period, the SRDP project will support the survey where possible, but allow the survey to develop independently as necessary. Once data processing and quality assurance operations have matured, SRDP will incorporate these processes into the overall SRDP operations model.

As with any change to the operational model, a component of community outreach, education, and support is necessary. At the observatory, these tasks are addressed by the User Support groups and the project will support them in their efforts to educate and support the user community on the qualities and limitations of science-ready products.

2.1.1 Excluded Scope

In defining project scope what is not part of the project scope is often as important as what is. Functionality and capabilities which will not be delivered by the SRDP project are discussed below.

Creation of Derived Products

The primary focus of the project is on calibration and imaging of interferometric data, and providing effective means for discovery and access. Throughout this document we will use the term derived products to refer to products generated by the processing of images, examples include but are not limited to: line lists, source lists, moment maps, and position-velocity curves. Creation of persistent derived products or other forms of analysis are not included in the project, as the successful generation of science quality images is a pre-requisite. Simple analysis (such as creation of moment maps) may be included through the Archive interface, but are not primary products of the SRDP project and will not be archived. Similarly, curating catalogs generated by other projects is a service of the SRDP project, but the actual generation of the catalogs is beyond the project scope. This decision may be revisited after successful achievement of the current scope, either in the context of an extension to the current project or as a separate project.

Guaranteed Sensitivity Scheduling

VLA scheduling is based on the allocation of array time. As the observatory offers science quality data products as part of the proposal process, making the switch from a guaranteed time to a guaranteed sensitivity approach might make sense. ALMA already uses this mode of operation and a uniform policy for NRAO and ALMA telescopes has advantages. However, guaranteed sensitivity based operations would require significant changes to the overall operation of the observatory. Given the magnitude of changes, the SRDP project will not seek to change this policy, and assumes the observatory will continue to operate in a guaranteed time mode for telescopes other than ALMA.

Pre-Archive Processing

Modification of the VLA or VLBA operational model to move data processing closer to data acquisition (i.e. processing prior to archiving of the raw data) in order to decrease latency, or otherwise modify the data flow prior to the initial archiving of the visibility data is not considered in the SRDP project. Similarly no modifications will be made to the visibility data prior to the initial archiving (e.g. on the fly calibration) by SRDP processes.

Analysis Interfaces

NRAO will continue to provide tools to enable analysis of the resulting science products. Development of new tools to enable analysis are not included in the baseline scope of the SRDP project. The CASA package currently fills this role, and will continue to do so for the immediate future. The need / desire for online data analysis tools will be evaluated as the community transitions to more use of observatory generated science product.

User Spaces

Many of the organizational requirements of the Archive Use case would be addressed by the concept of a "User Space" in the archive. The concept of User Spaces is similar to DropBox, optimized for astronomical data. This construct would provide a unified dashboard for all of the users activities, with displays for several views:

- Observations (links to Proposals, Scheduling Blocks or other observation specification, and the resulting Execution Blocks)
- Data Products and Queries
- Processing Jobs
- QA Interface
- Downloads

Usage information, quota policy, and history could all be supported within this framework. From the user perspective, the space concept could be used to allow sharing of results, collaboration on reductions, and if extended to large projects, a means of monitoring and tracking the progress of such projects.

Fully developing these objectives are out of scope for the current SRDP project, due to time and resource constraints. Portions of these capabilities will be required for the SRDP project, but will be restricted to the minimal set necessary for delivering science ready products.

One capability that must be implemented for SRDP is the concept of a user cache, a location to store temporary products while they are reviewed, or as inputs to a successive processing step.

2.2 Actors

2.2.1 Human Actors

Maximizing the scientific impact of the NRAO telescopes means that we must support multiple types of researchers. In analyzing the prospective users of SRDP, it is useful to classify them both as to how they interact with the observatory and their level of experience. Of course, the same individual may belong to multiple classes of users depending on their current objective.

The first way to categorize users is by how they interact with the observatory. These categories include:

Telescope Users: Telescope users enter the system through a well-defined proposal process, and the Observatory has well established processes for adjudicating the relative merits of the proposals and allocating telescope time based on this process. Telescope users design observations to address specific scientific questions and envision the data products that will allow them to address those concerns. The storage requirements for the data products can be estimated and, barring reprocessing, bounds on the computational requirements estimated.

Archive users: Archive users enter the system through the archive interface and may be anonymous. These users seek to re-use existing data (and products) to answer scientific questions, which may be unrelated to the initial science case. Because this type of archival research represents a potentially unbounded computational requirement, we will require any user requesting data products which require additional computation to register with **myNRAO**. Automated processes will in place to monitor usage per user and large or expensive requests will be referred for manual evaluation on a case-by-case basis.

Two groups of Archive Users can readily be identified:

1) PIs, for which SRDP came 'too late', i.e. proposers who have not gotten around reducing / analyzing their data.
2) General user who is interested in specific datasets.

Policies for the prioritization between these groups will be developed.

Large Projects: Large projects (both telescope and archive based) represent a limited number of significant investments both on the part of the PI and of the observatory. The SRDP project will seek to maximize the return on this investment. The special relationship between the Observatory and these projects is described in section 3.9.

Operations Staff: Members of the NRAO staff with functional effort allocated to supporting the operations of the SRDP processes. This category primarily comprises data analysts and scientific staff members supporting them in the execution of the processing workflows and quality assurance process.

A second way of categorizing users is by their experience level. We identify four categories of SRDP users:

General Public: As part of NRAO's commitment to education and public outreach SRDP must support the amateur astronomer, educator, or other member of the public interested in astronomy. This group of users is primarily interested in the produced images, which will be available for anonymous download. The additional requirement for this group is support for file formats they are familiar and prepared to work with, in addition to the scientifically focused file formats required by the research community.

Casual Users / Novice Astronomers: Some users may want to use SRDP to get an image quickly without having to know much about radio astronomy. These users may either download a pre-existing image, use the archive tool to produce an image from previously obtained data, or desire a simple path to obtaining new data from which an image can be built.

Future Power Users: This group of users might start at a basic level and then increase in sophistication (i.e., grad students, astronomers who have experience with data at other wavelengths and want to take advantage of the capabilities of ALMA/VLA, etc.). These users would typically start as the previous category but over time will interact more deeply with the data. The primary requirement of this group is that SRDP interfaces support a continuum of expertise, not just novice and expert modes.

Experienced Radio Astronomers: The current expert users of the NRAO telescopes may not require science-ready products to be able to use our facilities, but decreasing the data processing effort required before they can begin to do science facilitates the timely production of results. The project will seek to provide interfaces to engage these experts in the generation and quality assessment of science-ready products. Collaboration with these community experts allows the project to benefit from their expertise in reduction heuristics and quality assessment. Finally, the products produced by SRDP must earn the trust of this group for scientific use and uptake to become widespread.

2.2.2 Workflow Management System

Many of the use cases and workflows described in this document call for automatic behavior, such as automatically notifying a user or triggering a pipeline execution. In this document, these autonomous actions are allocated to the Workflow Management System. There may be multiple interacting sub-systems required to perform all of these roles in the final design but that implementation detail is left to be addressed by the Data Management and Software (DMS) team.

2.3 External Interfaces

Users will primarily interact with the SRDP system through software interfaces, even when the software is only mediating the communication between operations staff and the user. This section describes the six primary external interfaces of the SRDP system. Internal software interfaces are not described here, as they are part of the system architecture maintained by DMS.

2.3.1 Proposal Submission and Observation Planning

For telescope users, the initial contact with the SRDP project will be in the proposal submission process. This accomplished through the PST for the VLA and the OT phase 1 for ALMA. For simplicity, we refer to this as phase 1 of the proposal process throughout this document.

For the purposes of SRDP, this interface captures the scientific intent of the telescope user, and ensures that this intent is properly captured and passed to the downstream processing so that the correct products can be generated. The SRDP project will engage with the appropriate interfaces in ALMA and NRAO to generate specific requirements for the phase1 tool to enable this intent capture. Modification to the phase II tools (OPT and OT) and the online systems may be required to ensure the flow of data through to the post-processing stages. The complex nature of ALMA governance may delay or prevent inclusion of additional requirements in the ALMA OT.

2.3.2 Archive Interface

The archive interface is the user's primary means of finding, creating and accessing science-ready products. The archive provides data and product discovery capabilities, product inspection facilities, an interface through which custom products may requested (the Post Processing Interface or PPI), and product delivery mechanisms.

2.3.3 Weblog

The pipeline weblog is the primary record of the process that generated a set of science products. The hierarchical HTML tree provides information for quality assurance, provenance, and detailed inspection by experienced users. The weblog has already been well developed by the ALMA and VLA pipelines, but will be refined and augmented by the SRDP project to continue to improve utility and usability. The primary need for improvement is to reduce the complexity of information presented, assure that most typical users can find what they are looking for without the need for staff support from NRAO to use the weblog, and generally improve the organization of the user interface.

2.3.4 Helpdesk

The NRAO helpdesk is the conduit through which the user communicates with operations staff. Like the weblog, this is an existing interface that SRDP will reuse. To date communication with the user through the NRAO helpdesk has always been initiated by the user. The proposed use of the helpdesk in the context of SRDP changes this paradigm to allow communication to be initiated by the Observatory. Customization to simplify SRDP workflows, avoid cutting and pasting information to/from the Archive Tool and to allow automatic updating will be defined and implemented as part of the SRDP project.

2.3.5 Workflow Management Interface

The workflow management interface is used by operations staff to monitor and control the flow of SRDP generation through the workflow lifecycle. In particular, this interface focuses on mediating the quality assurance process, ensuring products do not become "stuck", and tracking delivery status.

2.4 Policy Implications

Adding science-ready data products to services offered by NRAO raises a set of questions that extend beyond the scope of the SRDP project. These policy questions are choices to be made regarding how the observatory is operated. We list these choices, our assumed policy, and the implications of this for the SRDP project.

2.4.1 Calibration Strategy

Users of NRAO instruments can be extremely creative in the design of their observations to meet some of the most exciting science goals. Observatory-derived SRDPs are not possible for these types of experiments, as the resources needed are not well-defined. Observatory-derived SRDPs can be delivered for an observation or project only if a proper observing set-up validated by the Observatory is used. Users will not be required to adhere to these requirements, but nor will the observatory be obliged to provide SRDPs for observations not meeting these requirements.

The intention of the telescope user to conform to Observatory calibration strategies so that SRDP products can be generated should be captured as part of the proposal process. By default, projects should be assumed to conform and sufficient information captured at this stage to allow the creation of the scheduling blocks and observing scripts by the observatory, to be reviewed and approved by the telescope user[*]. Special considerations such as high flux-density or bandpass calibration accuracy or the intention of the telescope user to observe in full polarization must be captured as part of the proposal and the observing strategy augmented to support the requested calibration. Proposers requiring non-conforming calibration strategies shall be able to "opt out" of the standard calibration as part of the proposal process. Justification for this decision should be captured as part of the proposal process and an alternative calibration strategy proposed. Modifications to the Proposal Submission Tool (PST) and Observation Preparation Tool (OPT) will be required to support this type of templated observation.

The ALMA Observatory defines as non-standard those observations which cannot be correctly processed using the standard pipeline. Non-standard ALMA projects will be supported on a best effort basis. Non-standard projects will not be automatically excluded, but the SRDP project may decide they are too difficult to support at the available resource levels.

2.4.2 Proprietary Period

The proprietary period of products created using SRDP processes and resources shall be tied to the most restrictive proprietary period of the underlying raw data. In other words, science ready data products will only be proprietary while the underlying raw data are proprietary – if the raw data are public then any data

[*] This is parallel to the model that ALMA is adopting for upcoming cycles.

products will be public. If multiple data sets are used to produce the product, the product is public only when all input data sets are public. This allows the SRDP project to accommodate the different proprietary data policies of ALMA and the other NRAO telescopes.

2.4.3 Archive Contents Policy

In determining what products should be curated in the archive there are three aspects to consider: completeness, utility, and cost. To maximize the completeness of the archive, as many products as possible should be ingested and made available. However, a complete archive that has very heterogeneous processing and quality assurance standards is not especially useful for the archival user, particularly if the provenance of the product is unknown. Archiving all possible products is also unbounded from a cost perspective.

The initial policy assumed by the SRDP project is that all science products in the archive must have undergone, and passed, a known quality assurance process. Those products generated through SRDP processes will have undergone a standard process and should be designated with an NRAO or ALMA QA approval as appropriate. Existing science products already in the archive that have not undergone this process will need to be designated with a suitable QA status.

Large projects going through the NRAO submission process are currently required to submit a data management plan as part of the observing proposal. This should include a quality assurance plan as well as definition of what products will be submitted to the NRAO archive and estimates of product size. Once generated (see the Large Projects use case), these products should be quality assured by the proposing group and submitted for ingestion into the NRAO data archive. In this case, the QA approval flag should cite the project as the authority for the quality assurance. For large projects where the existing SRDP QA processes are sufficient, users will be encouraged to opt into SRDP processing to decrease variance in the product quality.

In the baseline SRDP Plan, ingestion of user generated products* to the archive outside of the large proposal case is disallowed, both to constrain the scope of effort (and cost) required for SRDP operations and to ensure the quality of products is known. This decision may be revisited at a later point in the project as experience with science-ready products grows and the needs of the user community become clearer. One possible extension would be to allow users to upload products that have been published (and cite the relevant publication), effectively using the referee as a quality assurance process.

2.5 Computing Resource Management

The production of science-ready data products at the Observatory requires that NRAO have access to sufficient compute and temporary and long-term storage resources to address the needs of the community. Currently both the New Mexico Array Science Center (NMASC) and North American ALMA Science Center (NAASC) have sufficient resources for existing demand. Usage policies in place are designed to balance the competing demands from pipelines, large projects, NRAO staff, and the NRAO user community. A first order estimate for ngVLA anticipates image volume equivalent to the raw data volume, with a limit on larger full spectral resolution cubes (EVLA: 600 TBytes/year, ALMA: 300 TBytes/year). Calibration products are relatively small (0.1-1 GBytes per observation).

Production of science-ready data products will change the demand profile for the computing cluster and Lustre resources. While more data processing will be required overall, this will be partially countered by efficiency improvements as NRAO shifts from interactive manual operation to more automated pipeline and batch mode operation. NRAO's Data Management and Software (DMS) department and the SRDP project will work together to monitor computing resource usage and balance.

* User generated products are those created by a user independently of the SRDP framework. Products created using the SRDP framework at the request of a user will be ingested to the NRAO archive as described in the use cases in Section 3.

In order to manage resource allocation, a clear understanding of usage requirements and patterns must be developed. The SRDP systems must provide an accurate picture of usage patterns, and the ability to enforce quotas or other constraints. For any usage of the SRDP system, the user is expected to have a valid NRAO account, and to be properly authenticated through the ***myNRAO*** portal. The only exception to this is support for anonymous download of existing products from the archive.

Oversubscription of the NRAO computing resources is a risk for the SRDP project. Mitigation of this risk is primarily an implementation issue. However, we briefly discuss it here because these strategies will affect end-users. We consider two providers for support of SRDP processing on non-NRAO compute resources: Amazon Web Services (AWS) and NSF funded supercomputing facilities made available through the XSEDE project.

While existing pipelines for ALMA, VLA and VLASS have been demonstrated to work on AWS and XSEDE, further improvements must be made in reducing memory footprint and improving parallel processing functionality before either is a feasible option. Provided these technical issues are addressed, the following use cases can be supported. All cases are assumed to involve image processing; calibration pipelines have such low resource demand that external processing isn't necessary or practical.

Pipeline Processing
Standard pipelines may be executed on external facilities by NRAO as part of normal operations when demand exceeds existing internal resources. This processing model will consist of a special overflow queue on the existing clusters. Jobs will be submitted or migrated to this queue at which point remote processing would be triggered.

XSEDE resources are requested quarterly and then available for 12 months. Therefore, resource requests cover 3-15 months into the future, which makes them difficult to use effectively in demand overflow cases. AWS is attractive in this case since resources are only used on demand, although a premium over the raw hardware cost is paid for this flexibility. These are appropriate mitigations for spikes in demand; sustained levels of increased demand are best addressed by increasing the available resources at NRAO.

Special requirement processing
Some imaging cases may require resource profiles other than those supported by the NRAO cluster environment. For example, large scale imaging cases requiring greater than one terabyte of memory. In these cases, AWS could be more attractive than acquiring dedicated hardware.

Large project processing
Some large projects may take too long to process on local resources, or create operational issues for other projects. For instance, 1000 imaging jobs each taking 24 hours could be processed concurrently on AWS in 24 hours rather than weeks or months on local resources.

Charged User Processing
When processing load levels extend beyond NRAO's ability to respond, alternative processing may be identified to augment the internal computing resources. This could either be submission to AWS with some method of passing the charges back to the requesting user, or a block grant of XSEDE time which NRAO administers for the benefit of the community.

Charged user processing is not foreseen a primary mode of operation, but as a possible avenue for providing relief if computing resources become severely oversubscribed. Before adopting such a mechanism, a detailed policy would be drafted and discussed with the user community representatives. The most likely application would be for large blocks of processing which could exploit the increased parallelism of these resources.

Product Storage
In some cases, it may be attractive to temporarily store products in AWS or XSEDE to offset network

load or storage costs. This could be used as a temporary buffer against fluctuations in demand. For example, during data releases from the sky survey each product will trigger multiple downloads. An extra 100TB may be required for a month or two or for delivery of these highly anticipated products.

Data Archive

A copy of the ALMA or VLA data archive could be placed at an external facility rather than kept at a quiescent local mirror. This will simplify data transport if the external facility is also a processing facility. To date, this has not been cost effective, but DMS will continue to evaluate this option as time progresses.

2.6 Quality Assurance

The quality assurance process is a central component to the SRDP project. The objective of the process is to certify the reliability of the products, and provide a framework for users to understand the limitations of the products and therefore limit over- or mis-interpretation. The quality assurance process will necessarily evolve with the project, and a major initial effort to define the metrics to be used will be required. Within the duration of the SRDP project it is expected that QA will always require human inspection, although this may be a simple check that the process completed satisfactorily (such as in the restoration of calibrated data). Improvements to the QA metrics and pipeline heuristics will expedite the process, but the possibility of unanticipated features in the data requires that each product have a human review. Certain very uniform observing and processing modes (e.g. the Sky Survey) may be exceptions to this principle.

Quality assurance of calibration processes are relatively straight forward, if time consuming. ALMA has well established processes which can be modified and improved for the VLA, with the objective of reliable calibration that can be trusted as the initial step for an imaging process. Radio Frequency Interference is a particular challenge at the VLA frequencies, and additional techniques will need to be developed to address and quantify the impact of RFI on calibrations.

Image QA in the context of SRDP is significantly different from the standard imaging QA process for ALMA. ALMA standard imaging QA (QA2 in the ALMA quality assurance plan [**RD5**]) attempts to determine if the observations achieve the required sensitivity and can therefore meet the project science objectives, and if not then the project is returned to the observing queue. For the VLA the emphasis will be on determining the limits of the product and ensuring that unwanted artifacts are not present in the product. In cases where the user is working directly with the operations staff on a particular product, the user will be involved in the QA process as they are best able to determine if the product is suitable for their needs.

2.7 Summary of Use Cases

The following use cases are developed in more detail in Section 3. A concise summary of each use case is presented here to provide an overall perspective. The first four use cases provide the foundational capabilities that are the core of the SRDP project. Subsequent use cases support the accomplishment of these four, provide additional functionality or requirements for special use cases, serve to highlight specific requirements that are not explicit in the first four, or detail an important sub-process.

- **Standard Calibration:** Automatically generating science quality-assured calibration for supported observational modes.
- **Standard Imaging:** Automatically generating science quality assured images for supported observational modes. The objective is a homogeneous set of images in the archive to support telescope and archive users.
- **Optimized Imaging:** Generating specific science images as requested by a science user. Images will be quality assured (in conjunction with the user) and delivered to both the requestor and to the archive.
- **Archive Use:** Enabling data discovery, selection, creation of processing requests, and the delivery of data to users.

- **Restoration:** Returning one or more measurement sets to the calibrated and flagged state they were at the end of the standard calibration process. The calibrated measurement sets might be delivered to the PI directly or serve as the initial state for other processes.
- **Recalibration:** Repeating the calibration step, either with a different version of the supporting software tools, or with additional inputs from the user. Quality assured recalibration products are stored in the archive.
- **Combined Imaging:** Combining data taken from multiple configurations of a particular array or telescope (including the ALMA Total Power Array) to produce quality assured images with better flux recovery.
- **Time Critical Observations:** Accounting for time critical observations, including both triggered and target of opportunity observations. These observations modify the standard processes to decrease latency in product delivery.
- **Large Projects:** Large projects represent a significant investment both by the observatory and by the users. This use case focuses on maximizing the return to all stakeholders, and leveraging the special relationship between these projects and the observatory.
- **Curation and Reproducibility:** Assisting the NRAO user community by documenting reproducible data reduction processes.
- **Commissioning and Validation:** This case covers the special considerations that are necessary for NRAO staff to commission and validate the SRDP deliverables prior to their release.

3. Use Cases

3.1 Use Case: *Standard Calibration*

Providing standard science-quality calibration for most observations from NRAO telescopes is a prerequisite for nearly all of the subsequent use cases described in this document. This use case is also where the differences between the telescopes is most apparent. The ALMA observatory is responsible for providing standard calibration of all observations made using the ALMA telescope. To avoid duplication of effort, conflicting requirements, and redundant implementations, the SRDP project will not address the standard calibration process for ALMA data. We assume that the calibration has been generated, quality assured, and is available in the ALMA archive for application through the Restoration use case (see section 3.5). Although definition of the ALMA calibration process is outside of the SRDP project, the requirements of the ALMA process may be modified through representatives at the NAASC. Data processing effort managed by the SRPD project will be required to continue to meet the NAASC commitments for quality assurance of the ALMA products.

As discussed in section 2.4.1, the intention of projects to conform to SRDP standard calibration strategies will be captured as part of the proposal process. Projects which have elected non-standard calibration strategies will not trigger the standard calibration workflow. When a conforming observation is complete[*], and necessary meta-data for successful calibration are available, the observation shall be automatically triggered for calibration (as opposed to waiting for a request from the user). Auxiliary data such as calibrator fluxes, antenna positions, and known defective equipment shall be automatically considered as part of the calibration and should not require any additional effort on the part of the telescope user.

Calibrations should represent observatory recommended best practices at the time of execution (and thus will evolve over time), and should be congruent to calibrations which could be performed by an individual

[*] We assume that each execution will contain all required observations and intents to be calibrated independent of other observations. Observing efficiency improvements may be considered at a future time but are not part of the baseline SRDP project.

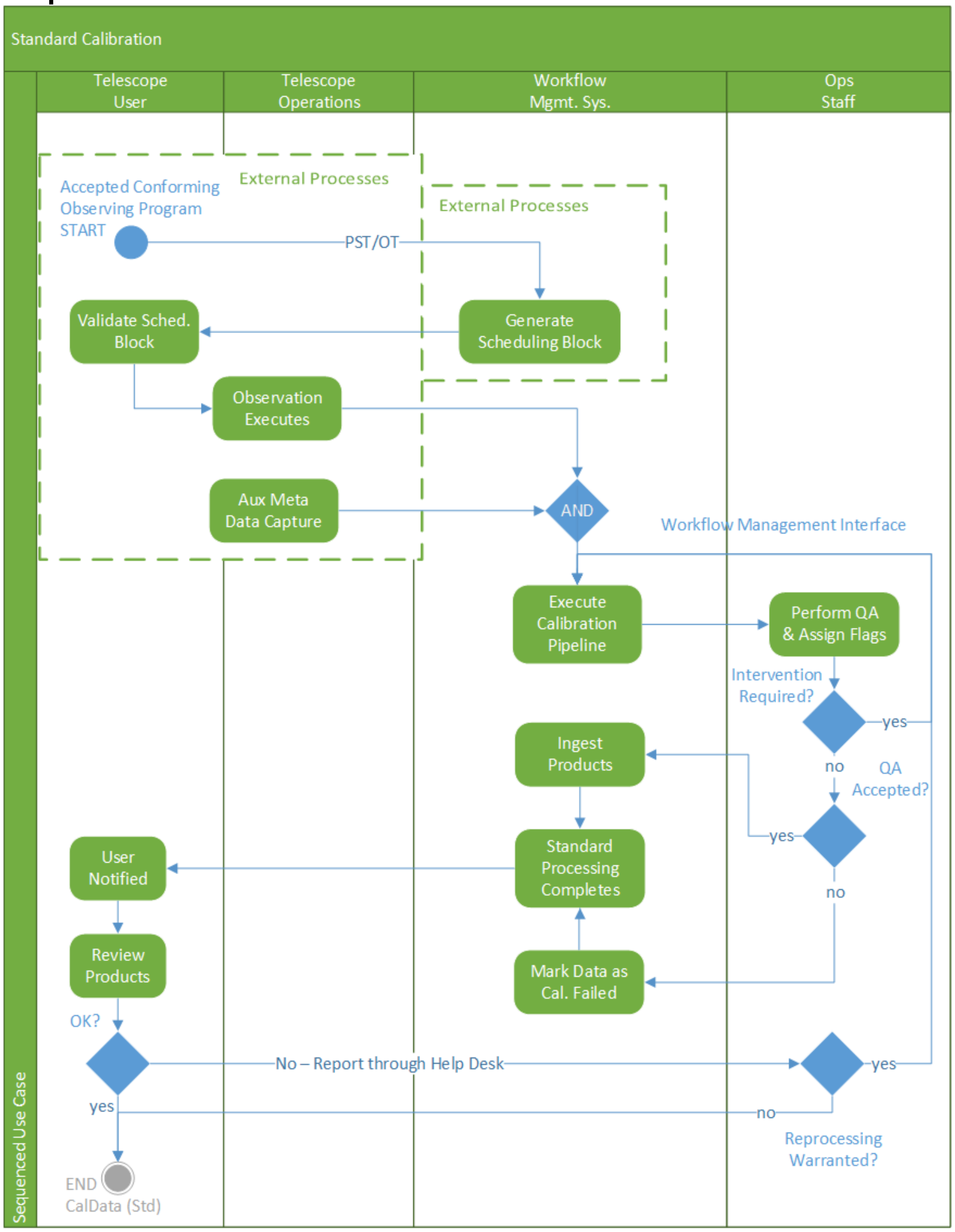
user. Every calibration shall be assessed for quality, and those projects for which the initial calibration are not judged to be of science quality should be identified for further intervention. Based on ALMA experience, the predominant intervention is the introduction of additional flags, either through automated or manual means. Any flags applied shall be captured in such a manner that the flags can be re-used by subsequent recalibrations (see section 3.6). Similar effort to maximize the utility of interventions in recalibration should be made as experience is gained.

Once a science-quality calibration has been generated for a particular observation, the calibration products, flagging information, and logs shall be ingested to the archive and the telescope user notified. The archive shall store sufficient meta-data to provide provenance for the calibrated products, and to promote identification of suspect products based on defects found at later times. Products for which a science quality calibration is not possible with a particular pipeline version shall be designated as such in the archive to prevent repeated attempts to calibrate such observations with said pipeline. Categories for failure shall be identified and metrics derived in order to allow the Observatory to address common failure modes. The latency between the completion of the observation and the delivery of products shall be measured. The goal is delivery of science quality products within 30 days of the completion of observations.

The user should be able to access the calibration and quality assessment results through the archive interface. The calibration record should be hierarchical in nature to support both summary and detailed views in order to support a wide range of expertise in the user community. To facilitate remote exploration of data, the record should make use of Data-Driven Documents (D3.js) or other similar visualization systems where possible. Quality metrics should be clearly identified and scores derived to simplify comprehension. Where possible physical quantities should be displayed as well as the normalized scores.

The helpdesk interface shall allow the PI to provide feedback on the calibration for a particular observation and request an improved calibration be performed based on a modified set of flags or other parameters specified by the PI. Processes to simplify this recalibration both for the PI and the Observatory should be in place, as well as a mechanism for designating the resulting calibration as the primary calibration for the observation.

3.1.1 Sequence of Events



Constraints

- None

Assumptions

- Submission tools (PST/OPT/OT) modified to allow proper capture of required information.

Pre-conditions

- Conforming observation defined in observing proposal.
- Auxiliary meta-data available from telescope operations.
- Special calibration criteria (e.g. high flux accuracy) are correctly captured in the observing proposal.

Post-conditions

- Standard calibration, logs, and quality assessment are stored in the Archive and available through the Archive Interface.

3.2 Use Case: *Standard Imaging*

The standard imaging process is already defined for the ALMA telescope, so the description below is applicable only to VLA observations. The standard imaging process will automatically be triggered for observations supported by SRDP once the standard calibration has passed quality assurance. By default, images for projects with multiple executions of the same scheduling block in the same configuration will be generated when calibrations for all of the executions are archived, or at the end of observing in that configuration. Projects for which image products are required from each execution, will be identified as part of the observing proposal and the trigger for the imaging pipelines adjusted accordingly.

The standard imaging use case is designed to populate the archive with consistent high-quality images that can be used for science research. For the telescope user, they provide a quick check of the calibration quality and default image. For many telescope users, the products may be used directly for science, although in some cases an optimized image (see section 3.3) may be necessary to fully achieve the science goals.

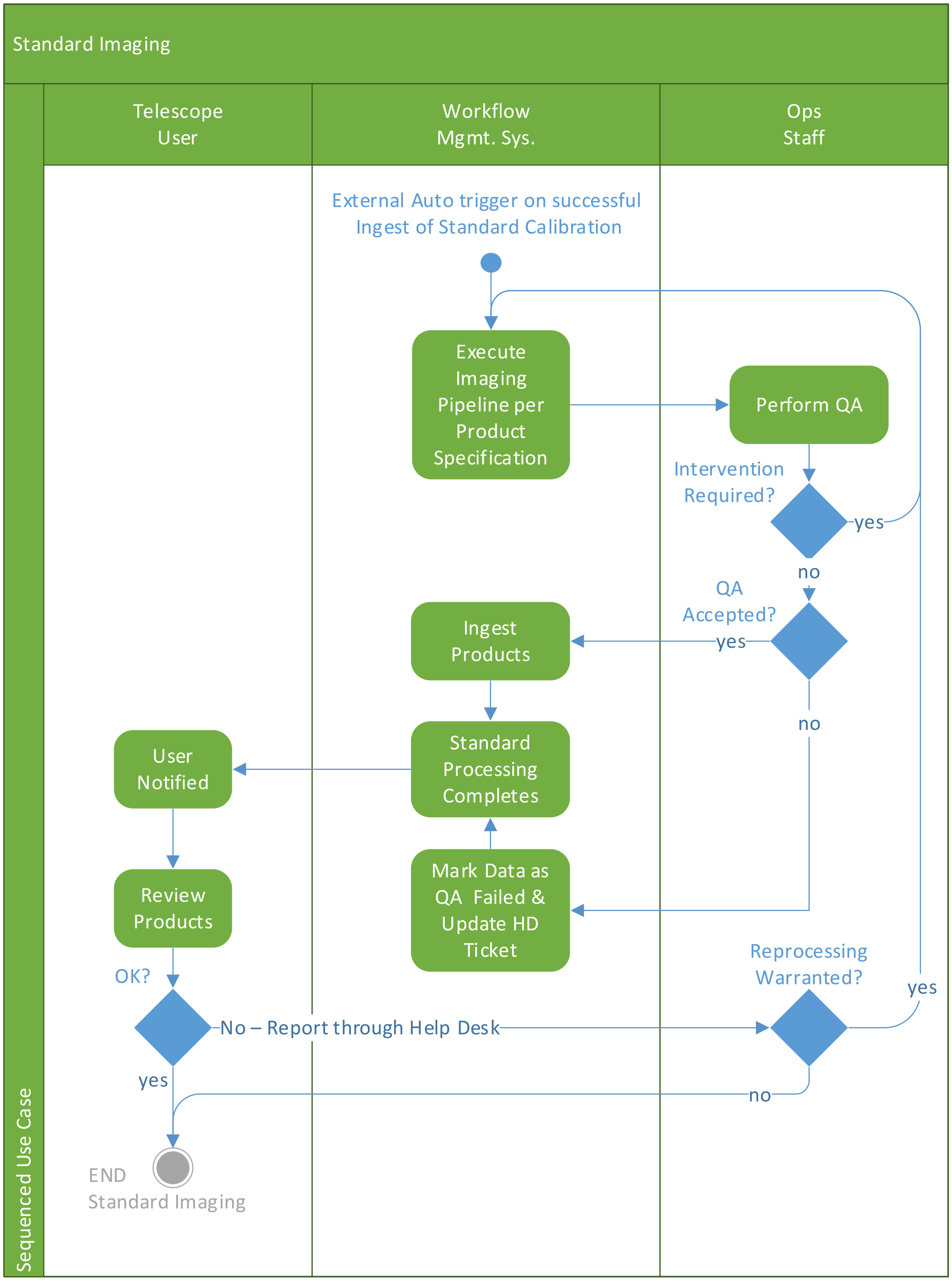
Archive users may find the standard products useful for science, as the images will have been quality assured and the flux scale will be well understood. The standard products may also be useful for archive users in assessing if an optimized image generated from the same data would be useful for their application.

The definition of standard image products must balance the requirements of the telescope user, the desire for a rich and homogenous archive, and resource constraints both in the generation and storage of products. For all projects, a full bandwidth Stokes I continuum image should be produced per receiver band, combining multiple pointings in a mosaic if specified by the project. For fractional bandwidths greater than a threshold value, spectral index maps should also be generated. For spectral imaging projects, cubes should be generated and archived at the spectral resolution specified by the telescope user, provided that the products do not exceed reasonable limits on size and computation resources. In cases for which the requested spectral cube is determined to be "too large" the user shall be informed at proposal time and allowed to refine the requested product (e.g. spectral range or resolution) to conform to size and computational limits.

Like the standard calibration process, most projects are expected to be supported by the SRDP project. However, for projects that cannot conform to the SRDP requirements, PIs will be required to opt out at the proposal submission stage with a brief description of why SRDP imaging is not appropriate for the project. For the majority of proposals for which SRDPs will be produced, sufficient information needs to be gathered at the proposal stage to capture the proposers' desired imaging product. These parameters should specify image characteristics (as opposed to processing instructions) and shall include the desired spatial and spectral resolution (for non-continuum projects), as well as whether multiple phase centers are to be imaged separately or are intended to be mosaicked.

The workflow for the standard imaging case is parallel to the standard calibration use case, using the automated workflow manager to trigger creation of the images based on specifications from the observing project. Operations staff will perform quality assurance on the products, and communicate with the telescope user through the helpdesk interface. Once implemented, the goal is to make standard SRDP images available to the telescope user within 30 days of the required data being acquired at the telescope.

3.2.1 Sequence of Events



3.2.2 Constraints

Assumptions

- **None**

Pre-conditions

- Standard calibration has completed successfully, and calibration is available in the archive.
- Image product characteristics were specified at proposal time and are available to the SRDP workflow.
- Spectral line projects are identified at proposal time and the desired spectral resolution is specified metadata.
- Non-conforming projects are explicitly identified by the PI and will not trigger the workflow.

Post-conditions

- A full bandwidth, Stokes I continuum image is archived for each observing band in the observation.
- For fractional bandwidths greater than a threshold value, spectral image maps are available in the archive.
- For spectral imaging project, cubes at the spectral resolution specified by the telescope user are available in the archive
- For polarization projects, Stokes Q and U images are available in the archive.[*]

3.3 Use Case: *Optimized Imaging*

A user wants a science-ready image of data identified through the NRAO archive search interface. At least one validated calibration is assumed to be available for the data (missing calibrations can be requested through the recalibration use case, and the workflow should start with those calibrated data located in a temporary area).

Through the archive interface, the user will specify the desired scientific properties of the image (field of view, spectral extent, spectral and spatial resolution, and polarization). Reasonable defaults should be presented to the user and invalid options hidden. In addition, imaging pipeline parameters shall be optionally specified through this interface. Parameters should be scientific in nature and not tied to a specific implementation of the imaging process (for instance, specifying a desired RMS is appropriate, specifying the number of clean iterations to be used is not).

The request should be automatically validated, including a check that the data are available, the request is well formed, and the user has permission to access the data. Additionally, a check for identical reductions shall be performed to ensure that duplicate images are not produced. If for any reason the request is deemed invalid, the reason should be displayed clearly through the interface and the user provided the opportunity to either modify the request or automatically transfer the issue to the associated helpdesk ticket. This helpdesk ticket should be marked for manual follow-up, and the process will wait for manual resolution by operations staff.

Once the optimized imaging request has been submitted, a NRAO helpdesk ticket should be automatically created to provide tracking and communication between the SRDP operations staff and the user. This ticket should be automatically populated with relevant the request information.

The workflow process shall begin by restoring the data (see section 3.5) to the calibrated state, using the appropriate version of CASA and pipeline. Alternatively, a custom calibration created through the recalibration workflow (section 3.6) may be used. The image product shall only be ingested to the Archive if the calibration used is known to be stored in the Archive. The automated pipeline should be used to

[*] Production of all products is subject to reasonable size and computation constraints.

produce images and auxiliary meta-data (such as quality assurance plots and the weblog). When complete, an operations staff member is notified that the products are ready for quality assurance.

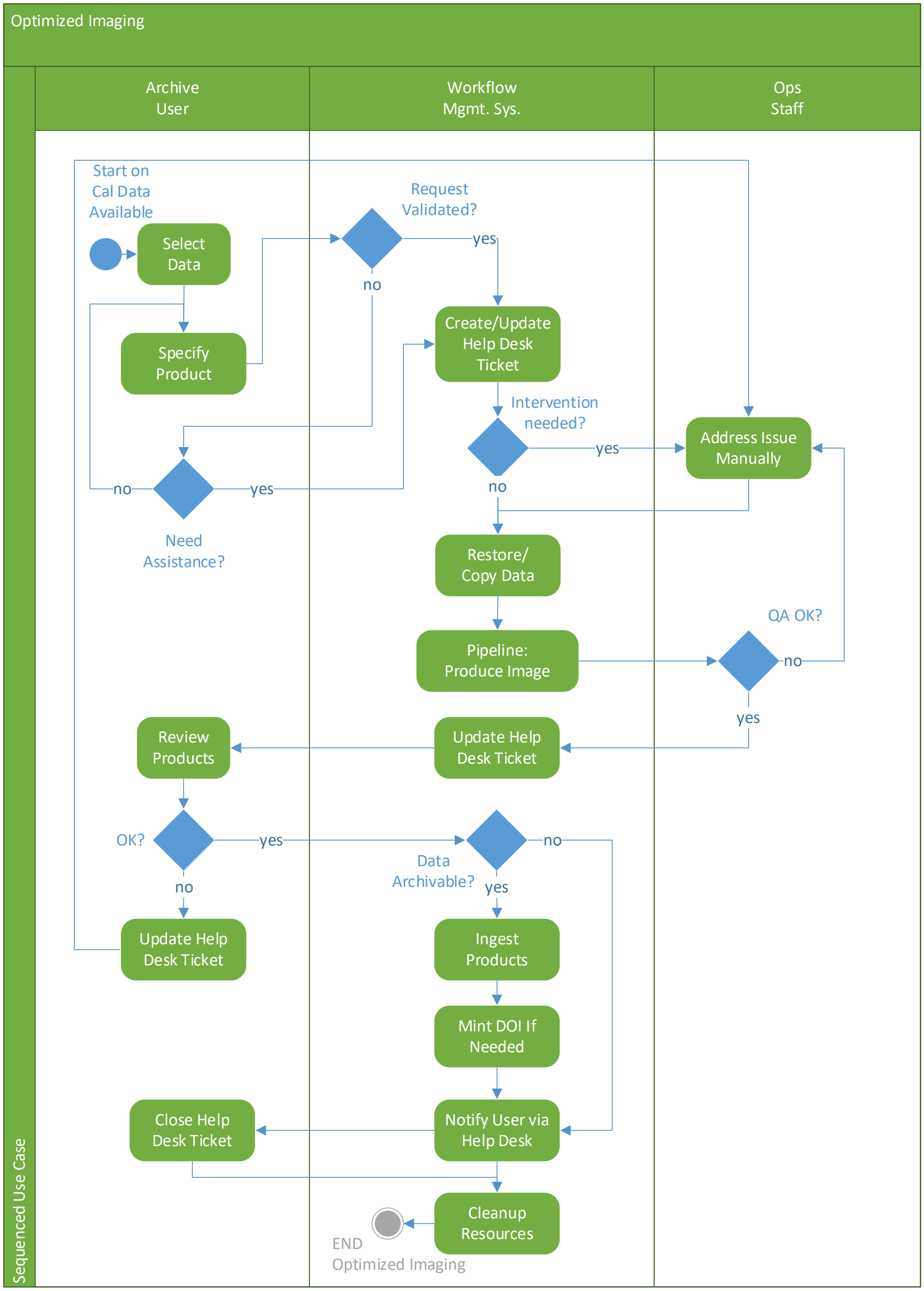
Quality assurance processes for optimized images shall maintain the same minimum level of quality as for the standard products. Any issues with the quality of the product images will be corrected by the operations staff member, in communications with the requesting user as necessary. When the requested image has passed quality assurance, the user is notified and the image and web-log are made available for inspection and download.

If the user is not satisfied with the product (for whatever reason), they can return to their request or helpdesk ticket through a provided link, modify as necessary and resubmit. A simple mechanism should be provided to ask for more assistance through a linked helpdesk ticket mechanism. Note that this iteration is potentially an open-ended resource commitment for the observatory, both in computing and staffing resources. Strategies to limit, or curtail the use of these facilities may become necessary, but it is premature to specify those requirements at this time. This risk is identified in the project risk register and technical mitigation of the computing hardware risk is discussed in section 2.4 above. The risk of staff oversubscription cannot be entirely mitigated through technical means. Efficiency improvements and partial automation of quality-assurance processes will provide some relief. The balance of offering capabilities to staff oversubscription will be managed as part of the SRDP project. If the user or staff member determines that a suitable image cannot be produced, this should be noted in the helpdesk ticket and the request canceled (removing it from the list of projects that are pending).

If the user is satisfied with the image a "validation button" will be provided to ingest the products into the archive (and optional creation of a DOI see 3.10), and the request closed. Ingestion of products will be subject to the same practical size and resource limits as standard products described in 3.2 An interface must be provided (probably as part of the workflow management interface) that allows for operations staff to identify requests which are pending, and follow up with the user to resolve the request. Depending on volume this checking and follow-up may need to be automated for common cases rather than relying on manual effort in each case.

Although true for all of the use cases, the Optimized Imaging case particularly could result in unbounded requirements for effort and compute capabilities. The capabilities for this use case will be released to the community in carefully designed stages, with small modifications permitted at first and increasing flexibility over time. In this way, both the operational processes and computational resources can be managed to ensure high quality of the products and process.

3.3.1 Sequence of Events



3.3.2 Constraints

Assumptions

- None

Pre-conditions

- Helpdesk tickets can be created and modified by the workflow system.
- At least one calibration exists for the selected data. If it is a non-standard calibration created through the recalibration workflow the calibrated MS must be available.

Post-conditions

- Image product is available to the user through the Archive Use case.
- The image is ingested to the Archive, provided full provenance information is available from the Archive, including the calibration information.
- The proprietary or public nature is determined by the nature of the underlying data.
- Helpdesk ticket is closed and resources used for processing have been released.

3.4 Use Case: *Archive Use*

This use case comprises many sub-cases which may be linked together in multiple ways. Presenting them together as a single use case emphasizes that there should be seamless integration in the tools, such that a user is presented a single view encompassing all of these use cases. The separate classes of Archive User identified in Sec. 2.2.1 shall be considered during detailed implementation of this use case.

Data Discovery

The user wishes to identify data objects in the NRAO collection that are useful for a particular objective. The archive interface presents a dynamic form with fields that may be used to filter the contents of the archive. The user selects relevant search criteria and is presented with the search results in a table. The search results table shows a set of default fields, but the user can configure which ones they want to see in the table; each registered user should be able to specify their "default view" and have this persist across searches. Similar options to sort the result on each column should be available to the user. For users interested in multiple searches a scriptable interface shall be provided, as shall the ability to export the search results to file in comma-separated value (CSV) or another file format.

Data Product Visualization

In addition to the limited information available in the tabular display, additional information and visualizations should be available through the archive interface. This should include:

- Image thumbnails displayed in context with other catalogs and survey results.
- For image cubes, quick look spectra should be shown.
- Provenance of the data product including links to the original data, other versions of the product as well as information on how the products were created, (processing job information, pipeline version, weblogs, etc.)
- Related publications, abstract for the project, etc.
- Online exploration of the data through a web enabled viewer (Such as CARTA or Aladin Lite)[*].

[*] Development of a web enabled viewer is outside the scope of the SRDP project, integration of an existing or emerging technology with the NRAO Archive is in scope.

The intention is to allow the user to explore data without needing to download large quantities of data, scientific analysis is not supported through this interface.

Data Selection

The user may select one or more data sets for retrieval, additional processing, or other type of follow up. Registered users may create a personal list of products that they want to investigate. These lists should be persisted across login sessions and multiple lists should be supported. Persistence can either be specified at the level of the query (in which case the result may change each time the query is executed) or at the level of the results (in which case the result is fixed). These catalogs are references to objects already stored in the archive, and do not point to temporary objects on disk.

The user shall be able to annotate and add tags for data products. In general, the tags are free-form, and only visible to the user that creates them. It is up to each user to decide how they want to use them. A special publically visible set of enumerated tags may be defined and used by NRAO. Large projects may similarly define public tags to assist in curating the products they produce. Examples of such tags would be a data release, or observation epoch identifier.

Data Processing

For each selected data product, a set of relevant processing options are presented. Options to begin other workflows described here (restore, re-calibration, optimized imaging) should be provided. A second class of lightweight product manipulation tasks such as generating a spatial or spectral cutout or providing a moment image shall also be provided.

Each of these jobs, once created, should provide the user an option to modify the input parameters and review the job prior to submission to the processing queue. The archive interface will provide status information for the user on each job, links to completed jobs, as well as the weblog for the job.

Quality assurance will be performed by an operations staff member. Additional user review of the products can be accommodated either through download of the data products or a temporary staging to the NRAO cluster. Once accepted, products meeting the requirements for archiving should be ingested to the archive. Ingestion of products will be subject to the same practical size and resource limits as standard products described in 3.2. To ensure the integrity of the product a checksum or other mechanism shall be used to ensure that the archived product matches the one produced by the processing both on ingest and on export.

Results of data processing may be temporarily cached, such as caching the results of a custom re-calibration prior to imaging the data. To prevent resource exhaustion, these results must be temporary and the automated system shall have the ability to automatically enforce the data retention policy. Warnings should be issued to the user two weeks prior to data removal with a reminder five days prior to the removal.

Data Delivery

Data products either generated by the data processing or otherwise selected through the interface can be bound together for delivery. Product download should allow individual products to be selected for download. Similar to the shopping cart on most web pages, one or more products can be added to the delivery "basket."

Several options shall be made available for delivery:

- a password protected URL that can be directly accessed
- a download manager capable of starting, pausing, and resuming download
- automated staging of data to the users' work area either in NRAO Socorro or Charlottesville.

Additional modes of delivery such as insertion into Amazon S3, or through the XSEDE frameworks, will be considered as experience and user demand dictate. The delivery process shall provide mechanism to ensure that data corruption through the delivery process is detected.

3.4.1 Constraints

Assumptions

- The archive interface provides a single point of access to all scientific data products available from the NRAO.
- All data delivery processes are mediated through the Archive Interface.

Pre-conditions

- A suitable online viewer is available to allow data visualization.

Post-conditions

- None

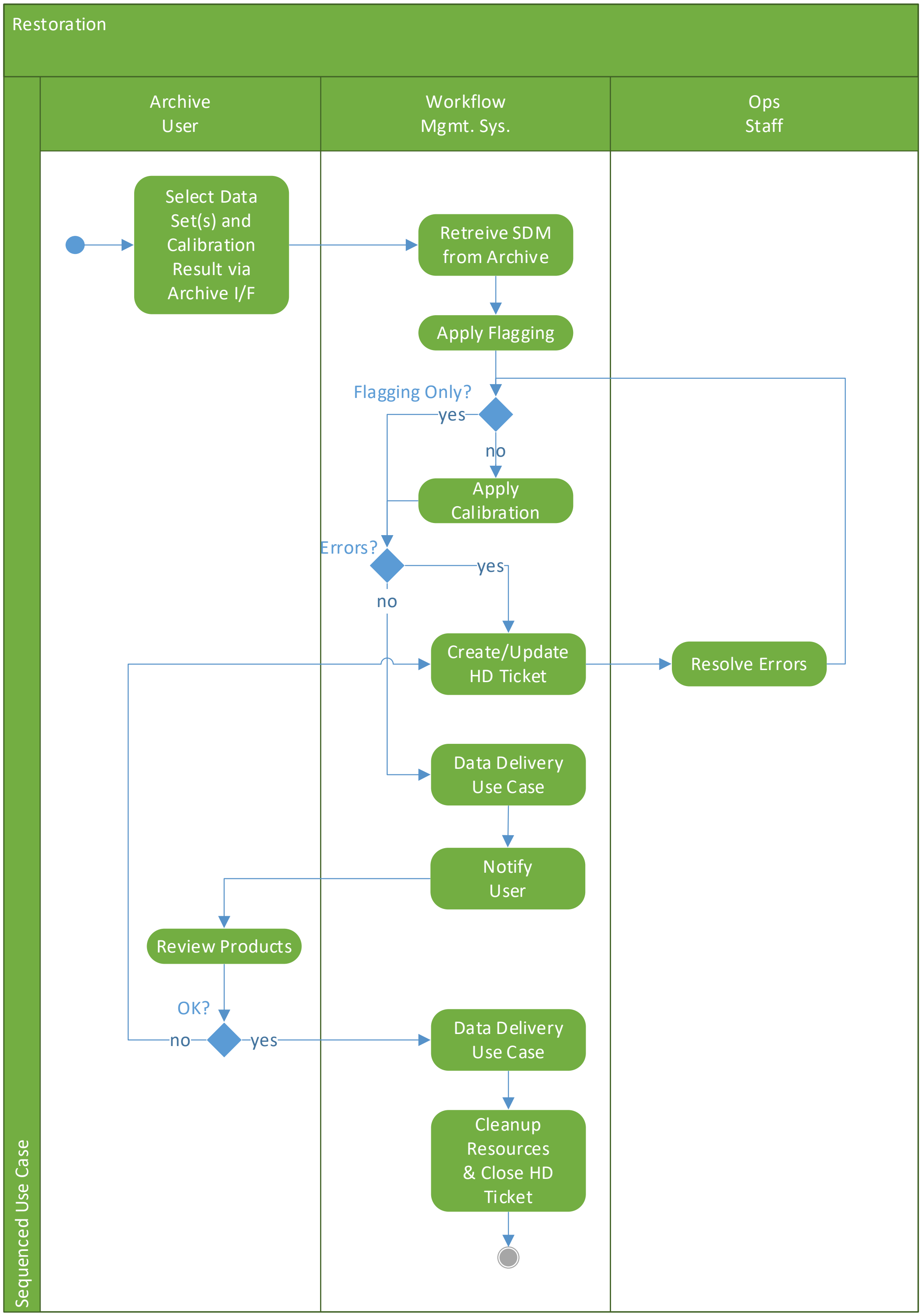
3.5 Use Case: *Restoration*

The restoration use case automatically returns raw data to a previously calibrated state. The archive user selects one or more data sets and an archived calibration result for that data set from the archive interface. Using the appropriate version of the pipeline and CASA, the raw data should be retrieved from the archive, the flagging state restored, and the calibration tables applied based on the instructions stored with the calibration results.

The restore use case has application both as a stand-alone process, and as the initial step of subsequent processing use cases. In the stand-alone case, an option to only apply the flags and not apply the calibration should be supported. In most cases, the stand-alone restore should be able to proceed without staff intervention (no quality assurance step is necessary), so no helpdesk ticket should be generated. If an error occurs during the processing a helpdesk ticket with the relevant information should be generated for staff troubleshooting and follow-up. The restored data should be delivered to the user through the standard data delivery use case.

When used as the initial step of other processing use cases, helpdesk tickets should be generated based on the parent use case and used for reporting of any erroneous conditions that occur during the restore process.

3.5.1 Sequence of Events



3.5.2 Constraints

Assumptions

- **None**

Pre-conditions

- One or more standard calibrations is available in the archive.

Post-conditions

- The restored/calibrated Measurement Set is available for use in other use cases or for delivery to the user.

3.6 Use Case: *Recalibration*

Recalibrating data sets, either individually or in a batch mode, will be required for a number of reasons. The originating user may be either a telescope user requesting recalibration of their proprietary data, an archive user requesting recalibration of public data, or an NRAO staff member performing the recalibration for maintenance purposes.

There are many reasons that the data may need to be recalibrated. The following specific goals have been selected to drive the described use case and derived requirements.

1. A user requests for part/all of a project to be re-calibrated in an older or newer version of the pipeline.
2. A user requests for a large number of projects to be re-calibrated in the old or new pipeline to have a uniform set of calibrations.
3. A user wants to modify the calibration strategy, for example using the phase calibrator for the flux scale.
4. A user wants to modify the flagging strategy used in the calibrations. This includes both adding additional flags to remove defective data undetected by the automated processing, or preventing the pipeline from applying flags (e.g. recovering spectral lines that the pipeline had flagged).
5. A staff member assisting a user needs to be able to reproduce their request and potentially submits an updated request for the user.

These objectives are not independent and reasonable combinations of the use cases shall be supported. The workflow for recalibration always starts with a user initiated request.

The user requests recalibration of one or more data sets either through the archive interface tool or through the helpdesk. A mechanism for the triggering of a recalibration based on search results should be provided. For each request, the user must specify:

- Sufficient identifying information for the data to be located in the archive.
- The pipeline version (including CASA versions if applicable) to be used.
- The desired calibrated products (i.e. calibration tables, calibrated measurement set, flagging information)
- *Optional*: Additional flagging specification
- *Optional*: Calibration strategy modifications
- *Optional*: Parameter modifications for the pipeline

If none of the optional parameters are specified, a check should be performed to determine if a valid calibration is already available in the archive, if so jump to the restore use case instead.

Once the request has been submitted, a NRAO helpdesk ticket should be automatically created to provide tracking and communication between the SRDP operations staff and the user. This ticket should be automatically populated with relevant the request information. The request should be automatically validated, including a check that the data is available, request is well formed, and user has permission to access the data. If for any reason the request is deemed invalid, the reason should be specified on the associated helpdesk ticket, helpdesk ticket marked for manual follow-up, and the process should wait for manual resolution by operations staff. If the job is large (either in number of data sets to be processed, or implied processing time), the request should be flagged for manual review by the SRDP operations staff.

Manual intervention, when required, will be performed by the operations staff. The staff member will work with the user to identify and resolve the issue and then resubmits the job for the user. At this point the process will re-enter the standard workflow.

The requests are scheduled for processing and status on the tracking ticket updated to reflect that the job is in the processing queue. Once the processing workflow completes the request is routed to operations staff for quality assurance. If no errors occurred during processing and no problems are detected in QA, the products are made available to the user through the delivery use case. A feedback mechanism through the helpdesk ticket should be provided for users to provide additional feedback, request additional changes, or accept the delivered results. The helpdesk ticket should not be closed until the products are accepted by the user, or it is determined that satisfactory calibrations are not possible with the data set. At this point, if the products are accepted by the user, then they should be stored in the archive. This calibration should only be placed in the archive (in addition to the previous versions) and made available to *other* users if only default parameters were specified, or if additional flags were specified to correct an issue not found during initial quality assurance.

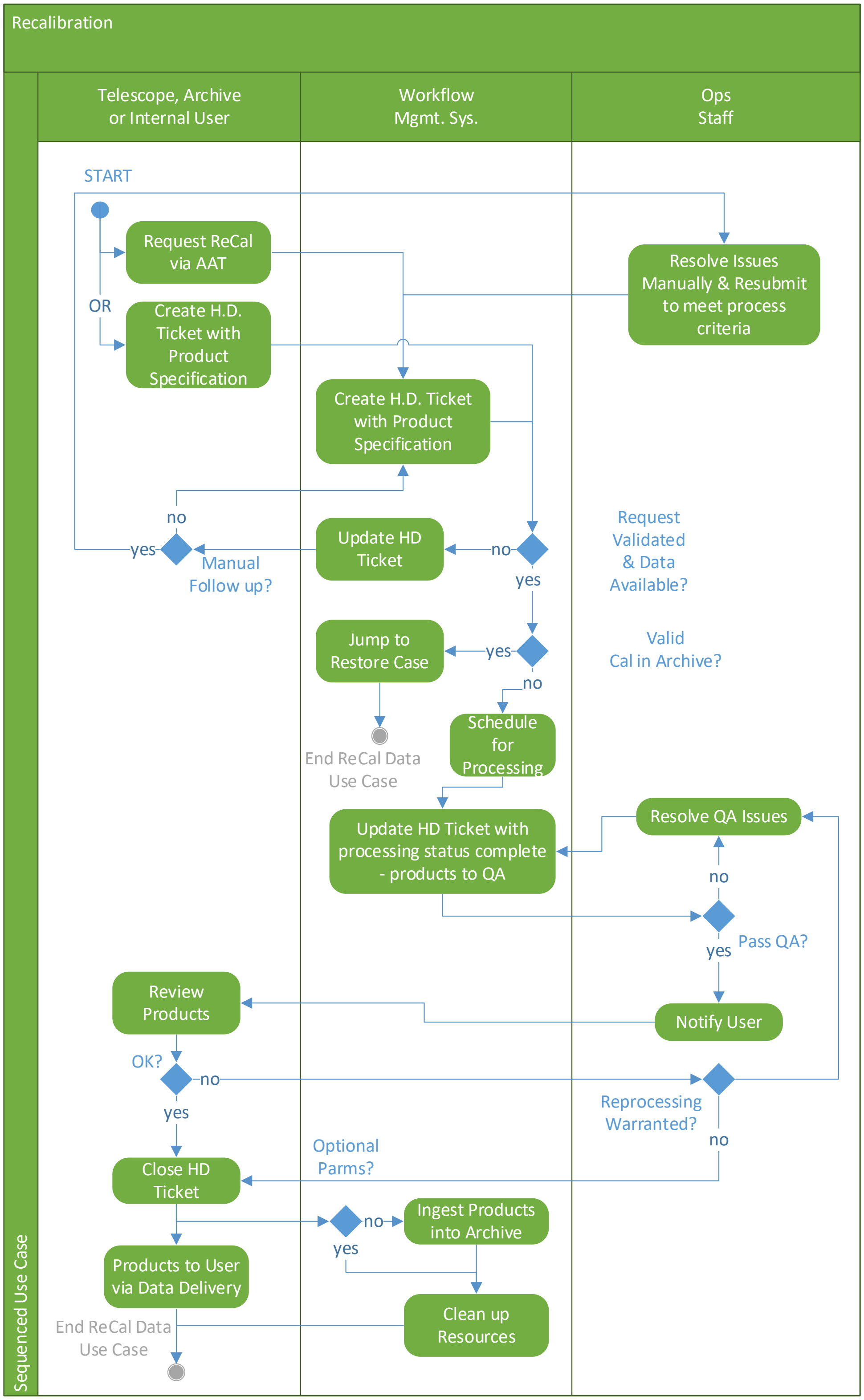
This calibration product should be made available to the user that created it as the basis for a subsequent imaging or other processing step, although again the subsequent products should not be ingested into the archive. A mechanism should be made available for the user to easily reproduce the same calibration result at a later date.

If errors occur during processing, or problems are detected as part of the QA process, operations staff assess the issue and in consultation with the user where appropriate either cancel the request, or resubmit it after appropriate modification.

Batch Recalibration

A special case of the recalibration use case is when a problem was identified with archive data. Staff members will need to re-process all affected datasets with an updated pipeline. Similar to the standard calibration, this is an observatory function. Since no external user-trigger is involved, a helpdesk ticket should not be created, the recalibration process shall be automatically triggered with a similar mechanism as for Standard Calibration in Sec 3.1 para 2. The recalibration process should be managed through the workflow system, including tracking of all affected observations and the managed submission of jobs to prevent overwhelming processing resources.. Affected calibrations need to be identified as no longer valid to prevent use of erroneous calibration.

3.6.1 Sequence of Events



3.6.2 Constraints

Assumptions

- **None**

Pre-conditions

- **None**

Post-conditions

- The resultant recalibration will be available to the user in their user cache, and may not be ingested in the archive.

3.7 Use Case: *Combined Imaging*

Multi-configuration imaging allows the astronomer to capture a wider-range of spatial scales in their image than possible from just a single array configuration. Simplifying this process for the user community is an important use case for the SRDP project. The goal is to combine data from different array configurations of either VLA or ALMA (including ALMA TP when available) to provide a final image. The case of adding total power data from other telescopes, or combining ALMA and VLA images is excluded from the scope of the SRDP project.

Two cases are considered below, the primary difference is if the observations were taken with the intention of combination or if an archival researcher is trying to use data taken for different purposes to derive a multi-resolution image.

Case 1 (Telescope User):

A telescope user proposes for a project that includes multiple configurations (VLA or ALMA) as well as possibly total power (ALMA). The proposal tool should automatically group the observations together and ensure that the spatial and spectral coordinates of the observation are consistent between the different epochs of observation. Total integration times for each configuration should be set according to observatory determined ratios.

As each configuration is completed the data should be calibrated and imaged independently using the resolution and pixel size most appropriate for the configuration, but with phase-center, field of view, and spectral axis of the common objective. This process should follow the standard and optimized imaging use cases discussed above.

When the single epoch calibration and imaging for all configurations are complete, the data from all configurations are imaged jointly, using the same spatial and spectral axes as for the individual configurations. The PI shall be able to specify an additional recalibration step to normalize flux scales, correct weighting issues, or otherwise normalize the data. The PI should be able to specify the same imaging parameters as in the standard and optimized imaging cases specified above.

If necessary, ALMA total power may be included at the end as a separate processing step. For standard products, current best practices for combining ALMA interferometric and total power data should be used (currently feather), while multiple options may be presented for optimized imaging cases. Diagnostic plots for the combined imaging are produced and included in the weblog.

Case 2 (Archive User):

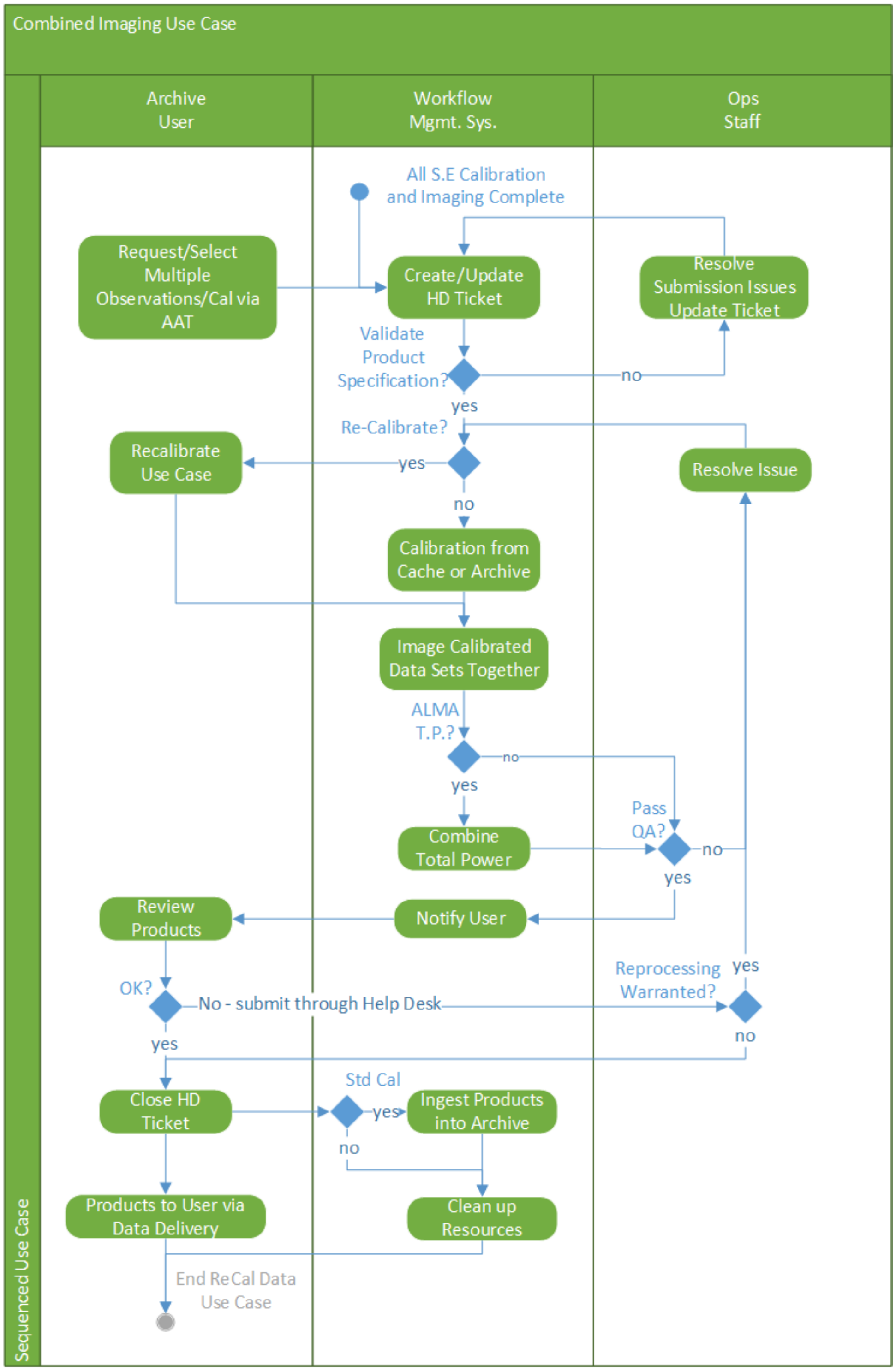
An archive user identifies multiple observations in the archive from separate projects that have similar enough spatial and spectral parameters that means that they could be combined with minor regridding. These observations could be from separate projects that could either be follow-up projects from the original research group, or from a different research group seeking to use for archival data for part of their project. It is assumed that calibrations for all of the observations exist or can be generated through a re-calibration process.

The user selects the observations to be grouped together from the archive, as well as the calibrations to be used. If necessary, the user may re-calibrate the data to e.g., ensure a consistent flux scale or version of CASA. Thus, input calibrated data sets may come both from the user's cache space and from data currently in the archive.

The calibrated data sets are imaged together using the imaging parameters specified by user. Unlike the PI use case, the spatial and spectral co-ordinates of the product cannot be deduced from the parent project and must be explicitly set by the user. The selected parameters must be suitable for all data sets, and should be validated both for applicability and to ensure that the implied requested re-gridding is within tolerance. For example, the channel width must not be smaller than that of the coarsest spectral resolution data.

In both Case 1 and Case 2 the resulting image shall be ingested into the archive provided that the calibration used on the input data sets is available from the archive.

3.7.1 Sequence of Events



3.7.2 Constraints

Assumptions

- None

Pre-conditions

- Multiple configurations of data with quality assured calibrations are available in the archive.

Post-conditions

- None.

3.8 Use Case: Time Critical Observations

This use case leverages the capabilities of the SRDP infrastructure to decrease the latency between observation and delivery of scientifically useful products. For these cases, speed may often be more important than achieving the highest possible the quality of the products. This emphasis on early access to results modifies the workflow from the standard calibration and imaging cases discussed above. The objective is that the total latency to delivery be dominated by the required processing time and that the non-processing aspects of the workflow add no more than 6 hours to the total latency.

This use case focuses two somewhat separate, but related, observation types:

- triggered observations that were submitted at regular call for proposal, and
- target of opportunity observations submitted through a director's discretionary time request

This use case may also be applicable in other circumstances, such as coordinated multi-instrument campaigns or rapid cadence regular observing where results from past epochs influence the schedules of subsequent epochs.

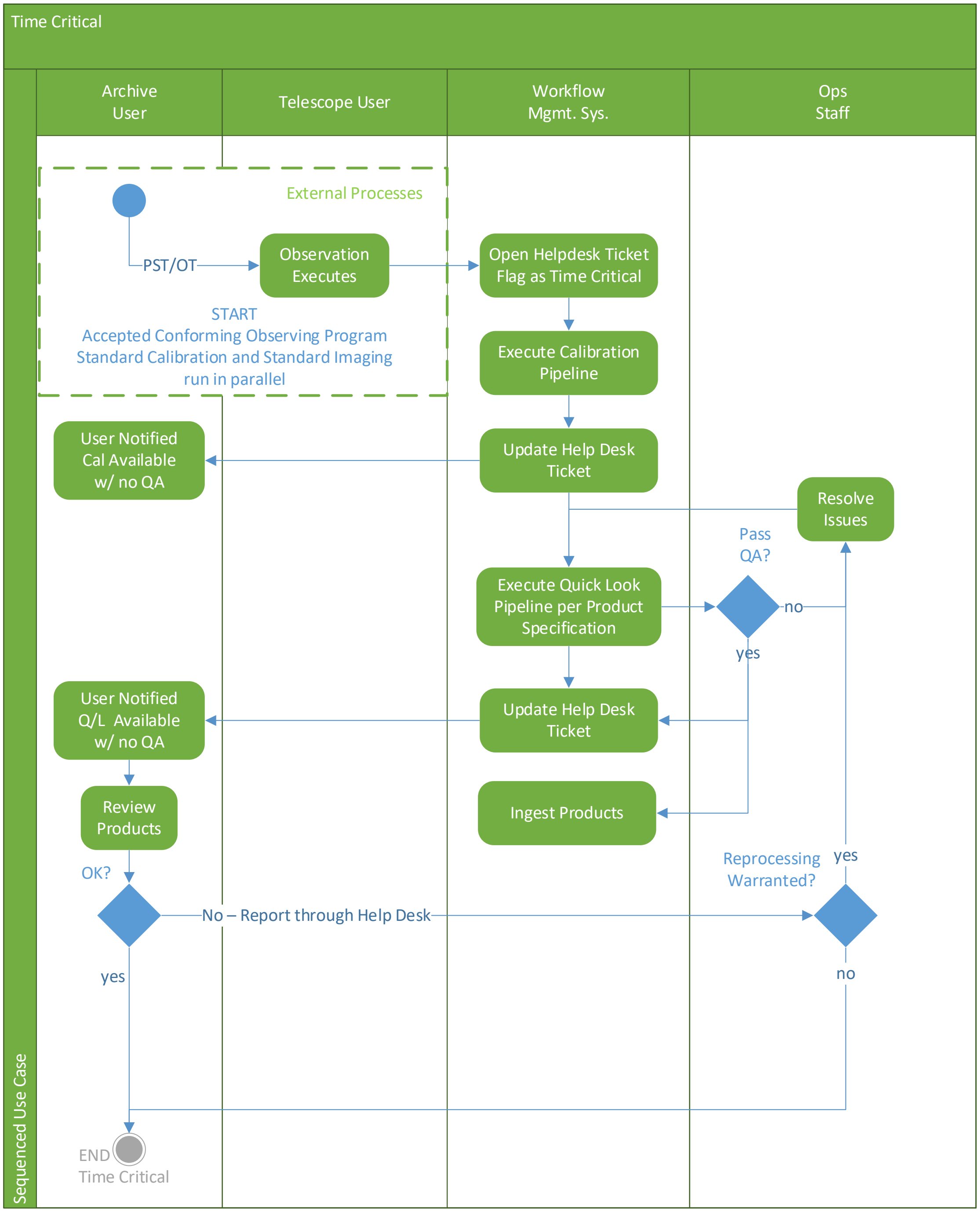
This use case shall be identified during the proposal submission process. The telescope user will indicate that the observation is time critical. This flag will persist throughout the lifecycle of the project and will be available to the data processing subsystems. The user should also be able to specify which data products should be treated as time critical: calibrated visibilities, quick-look images, or science-ready images. As with the standard calibration and imaging use cases, for SRDP products to be generated the user must conform with standard observing templates, and specify the characteristics of the desired imaging products. The quick-look images are a lightly cleaned imaging process optimized for speed rather than maximum quality.

Because of the additional scheduling constraints of these projects, they may execute at sub-optimum times. The SRDP system should not wait until auxiliary information such as calibrator fluxes or antenna positions has been updated prior to beginning reduction. For this reason, the standard calibration and imaging use cases should be invoked for these projects as well. In this case, both the clearly identified rapid reduction, and the later improved reduction should be archived.

Processing should begin as soon as data is available, and may be triggered based on elapsed time from the first observation rather than the completion of the data acquisition. This may require pre-empting the SRDP processing queue or making additional resources available specifically for this purpose. The PI should be notified immediately when calibration or imaging products are available, with specific notice that the products have not been quality assured. In cases of reduction failure, a high priority notification to operations must be made so that appropriate manual mitigation can be done. Note that this may occur outside of normal business hours.

Because of the rapid pace of these projects, transfer of visibilities or full data cubes may be untenable. As discussed in the data delivery use case, data assessment through the weblog should be supported as should remote viewing or transfer of image subsets.

3.8.1 Sequence of Events



3.8.2 Constraints

Assumptions

- None

Pre-conditions

- Project was identified as time critical during proposal process. Desired quick look products have been specified.

Post-conditions

- This use case may result in data that may not meet the quality required for the archive, increasing the requirements for the user cache.

3.9 Use Case: *Large Projects*

Large projects are defined as a class of observing programs that require significant time on the telescope (>200 hours for the VLA and >50 hours for ALMA). SRDP provides an opportunity to increase the return from these projects to the community, while assisting these expert teams in achieving the objectives of their project. Large projects can have a wide variety of levels of observatory involvement, ranging from mostly independent projects which take the raw data from the archive to process elsewhere, to projects like VLASS, where nearly all of the primary processing is done by the observatory.

All large projects are required as part of the proposal process to describe the data management plan for the data generated by the proposal, as well as the data release policy. Descriptions of the data products, quality assurance criteria and their approximate size should be included in future proposals. Large project teams will be encouraged to work with the SRDP project to maximize the scientific return to the community. The solicitation for large proposals and supporting documentation will provide supporting information.

The objective of the SRDP project is to support large projects throughout their execution. Because these projects are so varied, no single description of the use case addresses the needs of all large projects. Services described below are options that projects may avail themselves of and are not intended to constrain the project teams.

Data Acquisition and Workflow

Very large projects may require additional, specialized structures within the observing project, such as an epoch's worth of observations, or regions of sky. The SRDP project will work with each project to attempt to capture this structure in the archive to make provenance of the eventual products more traceable. This may include a specialized project specific "user cache" in the archive interface if requested by the project. These structures must be additional layers or views on the existing project structures to ensure that data is discoverable through the non-specialized archive interfaces as well.

Data Processing and Quality Assurance

Projects may opt to use the standard calibration pipeline for the calibration of the data, in which the standard calibration workflow (Section 3.1) should be followed. Some projects, like VLASS, will wish to develop custom or modified pipelines to process the data and be directly involved in the quality assurance process.

In a similar way, some projects may opt out of the SRDP imaging process as they require custom imaging or other post-calibration processing. If the standard imaging pipeline is sufficient the project may opt to follow the standard imaging workflow. Products produced by large projects may contain meta-data only meaningful in the context of the project, or even products that are not usually produced by the standard SRDP workflows. The project team will need to be involved in the QA process and ultimately is responsible for the scientific integrity of the products.

Using NRAO computing resources for the processing of the large project data may be considered provided that the required computing resources do not exceed those available (including prior commitments). Including this request as part of a large proposal, or allowing a proposal that only includes processing support and no related observing (for instance for a large scale re-processing of archive data), shall be considered as part of the SRDP project, once reliable estimates of computing load from other use cases have been developed.

Data Release

An important obligation of large projects is release of the data products back to the astronomical community. SRDP can facilitate this by hosting reasonable volumes of data products for the project. The large project may deliver a set of data products with at least meta-data conforming to a standard set defined by the SRDP project. Other project specific meta-data may also be specified and will be stored in the archive. This data will be marked as having received QA from the project team. SRDP may work with the project to produce a "project interface" to the archive. This is a dedicated search interface, that allows searching on the project meta-data as well as on the standard meta-data. This service may also be used by the project to describe the data, link to relevant publications, or otherwise provide branding and context for the results.

Products produced by the project may include standard types of products such as images and catalogs, but may also include other products. Decisions about which types of products will be archived are made by NRAO in consultation with the project team.

Commensal Projects

Commensal observing, parallel reduction of observations for different science objectives (such as transient science or simultaneous low frequency observing) increase the scientific return from a set of observations. Commensal projects are not subject to the usual proposal process, and thus are slightly different from Large Projects. Facilitating commensal projects in releasing products to the community is within the scope of the SRDP project. Future projects should identify the products and the release process as part of the negotiations with NRAO as the project is initiated. For existing projects the SRDP project will work with the project to identify and ingest appropriate products.

3.9.1 Constraints

Assumption

- Available resources (staff and computing) are available to support the large project request.

Pre-conditions

- The large project proposal process has been augmented for definition and review of data management and data products.
- Quality assurance standards of the large project are well described and available to the general archive user.

Post-Conditions

- Long term maintenance of the products from Large Projects will require small amounts of extra effort from the operations staff.

3.10 Use Case: Curation and Reproducibility

The reproducibility of results and open data standards are a major topic of discussion in scientific research today. The production of science-ready data products is an opportunity to assist the radio astronomy community in conforming to evolving standards in data provenance and process accountability. The structure of SRDP means that the observatory and any interested researcher, will have the full history of the processing done in producing a particular product and the means to reproduce the result if necessary.

Two deliverables are required from the SRDP project to enable users to be able to reference this information in reporting their findings based on observatory delivered science products.

First, the processing performed by CASA and the pipeline must be described in a publically accessible, and preferably referenceable location. Secondly, individual data products, and the processing history, must have permanent data locators to allow citation in publications.

Standards for the use of Digital Object Identifiers (DOIs) as persistent identifiers for astronomical data sets are being developed in the community, and the SRDP project should conform to best practices as they emerge.

The DataCite organization is likely to be the choice for the creation of DOIs for astronomical data. In order to create DataCite DOIs, NRAO must either become a member of the DataCite consortium or join the consortium as a partner in a collaboration.

The intention is that a unique DOI be produced upon request which provides a persistent identifier for the data product(s) and the provenance of that product, and that both will be directly accessible.

3.10.1 Constraints

Assumptions

- Standards for astronomical DOIs are still under development and could impact the delivery of this capability.
- Access to the DataCite consortium can be secured

Pre-conditions

- Publications describing the CASA package and Pipeline are available.

Post-conditions

- A "landing page" corresponding to the published URL must be maintained in perpetuity by the observatory.

3.11 Use Case: Commissioning and Validation

Throughout the SRDP project, the heuristics and operations teams will need to be able to test, commission, and validate parts of the system, or the entire system, prior to release for general use. The integrated nature of the system being developed could, if not carefully considered, preclude effective testing, or force the testing onto the production system.

The primary method of assuring the testability is the development of a written test plan as part of the development of each Level 1 functional requirement. However, there are several systemic capabilities that are required in support of commissioning.

1. It must be possible to execute SRDP workflows with candidate versions of the software. The products generated by this software must not be exposed as SRDP products in the standard data discovery interfaces.
2. It must be possible to execute portions of the SRDP workflows to optimize testing. For instance, avoiding the extraction of data from the archive and calibration each time that an imaging workflow is tested.
3. Due to long running jobs, it will not be possible to bring the system to a fully quiescent state prior to version changes, or updates. It must be possible to modify the system without losing the current execution state, or in such a way that the state information can be recaptured.
4. The execution environment may need to be modified, for example using a non-standard destination directory to accumulate outputs from a regression testing run.

If a duplicate or test system is used to fulfill some or all of the above requirements the test system must be identical to the production system in order to avoid unanticipated deployment issues.

To assist in system testing it would be beneficial if at any point where the user can specify calibration or imaging parameters, they can also request the scripts for download rather than actual processing. These scripts can then be used to perform testing in non-production environments. Some additional meta-data such as the paths to data and working area may need to be captured to allow complete scripts to be produced and not require manual editing by the user before execution.

In addition to the testing aspects for NRAO staff, the ability to download processing scripts may provide functionality for users wanting to utilize the SRDP capability at non-NRAO facilities. Because the data products produced this way will not go through standard NRAO QA, they will generally not be ingested to the NRAO archive.

3.11.1 Constraints

Assumptions

- None

Pre-conditions

- None

Post-conditions

- None

4. Overall System Description

The individual use cases enumerated in Section 3 describe various data flows through the system. This section looks at the SRDP system as a whole, defines the relations between the use cases, and provides a roadmap for the capability development. The view presented is not intended as a system architecture but rather to put the use cases in context both with each other and with the external system.

4.1 Products

Throughout the use cases, various inputs and outputs are implicitly identified, the following list calls out each product and provides a definition.

Archive Product: Any object stored in the Archive that can be accessed through the Archive Interface.

Auxiliary Meta-data: Additional meta-data from telescope operations used for calibration. Examples include: gain curves, antenna positions, and calibrator fluxes.

Calibrated Measurement Set: CASA working data format that has had instrumental and environmental effects removed. The Calibrated Measurement set can either be created through a calibration or restore process.

Combined Image: An image that uses data from multiple configurations to broaden the range of spatial emission scales captured in the image.

Custom Calibration: A calibration performed at the request of the user that does not conform to the SRDP standards, either due to flagging, intent modification, or other reason. The products are the calibration tables, flag information, and sufficient information to restore the science data model to a calibrated measurement set. Custom Calibrations are not stored in the Archive or made available to other users.

Digital Object Identifier: A unique and persistent string of characters used to identify one or more products in the Archive.

Log Files: CASA or other processing logs useful for detailed diagnostics or historical investigations.

Observing Project: The record of what the Telescope User intends to do. It encapsulates the proposal as well as the scheduling blocks (observing program).

Quality Metrics: Standard quantitative assessment of the product quality. These will be stored with the products in the archive. The intention is to give the user a measure of the Observatory's confidence in the product.

Quick Look Calibration: A calibration (calibration tables, flag information, and information required to restore the data) that has been optimized to minimize latency. Perfunctory QA is performed, but science quality is not guaranteed.

Quick Look Image: An image that has been lightly cleaned, but is likely not science quality. Generated for time critical observations to minimize latency.

Science Data Model: Data format currently stored for both ALMA and VLA telescopes; uncalibrated visibility data and associated meta-data to describe the data.

SRDP Scripts: A set of inputs which will drive the SRDP machinery to execute a workflow on a particular data set and produce a specified set of products.

Standard Calibration: The results of a calibration process that conforms to the standard processing recipe. The product contains the calibration tables, flag information, and sufficient information to restore the science data model to a calibrated measurement set of known quality.

Standard Image: Image produced and archived automatically for all conforming projects (with minimal input from the Telescope User other than in specifying the observing parameters). The archive of standard images will comprise an extensive homogeneous scientific resource.

Weblog: Set of hierarchical web pages designed to provide increasingly detailed information about a processing execution. This is the primary communication from the pipeline to the operations staff and Archive / Telescope User.

4.1 SRDP System Overview

Table 1 provides a summary of each of the use cases identified in Section 3, summarizing the inputs, outputs, and storage locations for each type of output. Many of the use cases have multiple sub-cases defined below them, each sub-case is presented.

Use Case	Inputs	Outputs	Storage Location
Standard Calibration	Science Data Model Aux. Meta-data Observing Project	Standard Calibration Quality Metrics Log Files Weblog	Archive
		Calibrated Measurement Set	User Cache
Standard Imaging	Science Data Model Standard Calibration Observing Project	Standard Image Quality Metrics Log Files Weblog	Archive
Optimized Imaging	Science Data Model Standard Calibration	Optimized Image Quality Metrics Log Files Weblog	Archive
	Science Data Model Custom Calibration	Optimized Image Quality Metrics Log Files Weblog	User Cache
Restore	Science Data Model Standard Calibration	Calibrated Measurement Set Log Files Weblog	User Cache
	Science Data Model Custom Calibration	Calibrated Measurement Set Log Files Weblog	User Cache

Use Case	Inputs	Outputs	Storage Location
Recalibration	Science Data Model Aux. Meta-data	Standard Calibration Quality Metrics Log Files Weblog	Archive
	Science Data Model Aux. Meta-data	Custom Calibration Quality Metrics Log Files Weblog	User Cache
Time Critical	Science Data Model Observing Project	Quick Look Calibration Quick Look Images Log Files Weblog	Archive
Combined Imaging	Science Data Model Standard Calibration Observing Project	Combined Resolution Image Quality Metrics Log Files Weblog	Archive
	Science Data Model Custom Calibration	Combined Resolution Image Quality Metrics Log Files Weblog	User Cache
Curation and Reproducibility	Archive Data Products	Digital Object Identifier	DataCite
		Archive Landing Page	Archive

Table 1 Summary of SRDP use cases, inputs and outputs.

The Archive Use Case is omitted from Table 1 because it describes the interface to the archive and not the production of new products. The specifics of the other omitted use case, Large Projects, will be largely defined on a case by case basis by the observing project and SRDP staff, and will have significant variation between projects. In the table, the designation of User Cache is used to designate products available to the user that requested them, but not permanently ingested to the archive. It is not intended to designate any particular implementation.

Figure 1 shows the data flow diagram for the SRDP system at the use case level. The general flow is from the top left to the lower right. Use cases are either triggered from the automated workflow, or instigated by users through the archive interface.

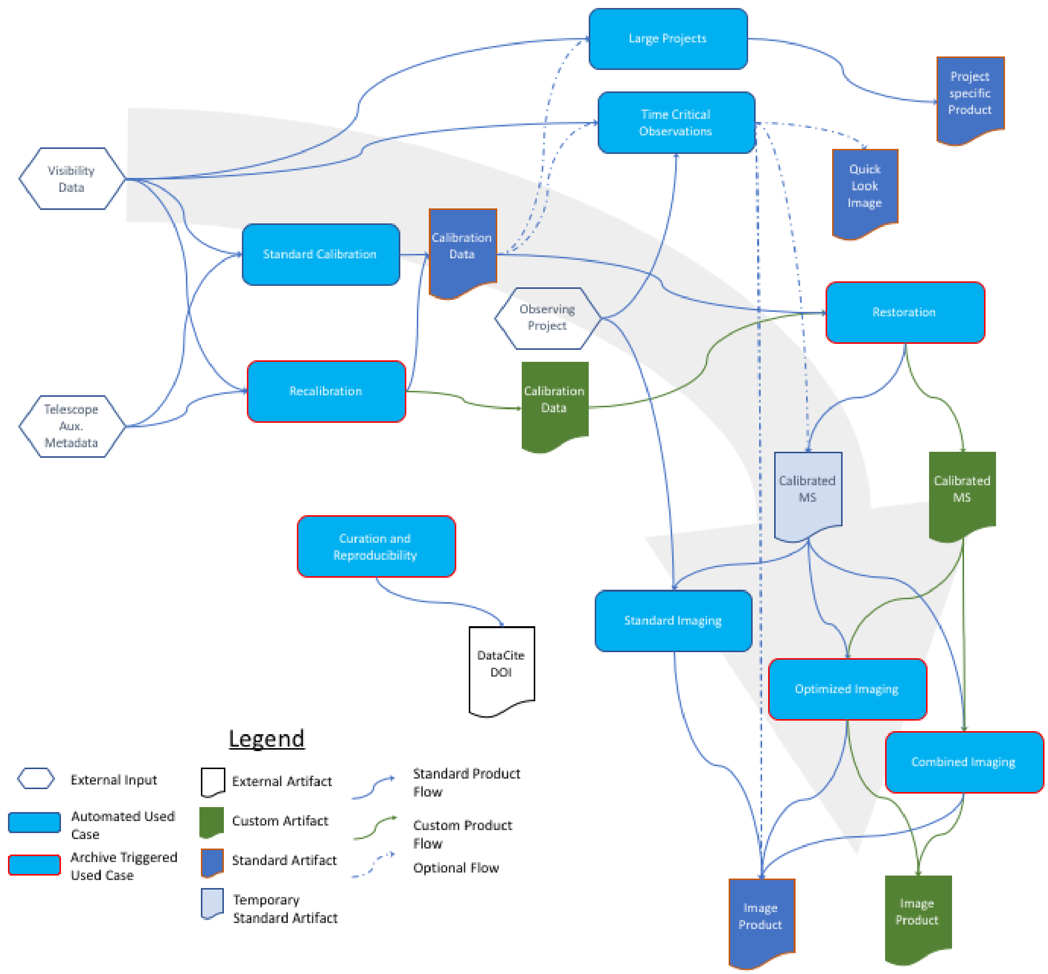


Figure 1: Data flow diagram illustrating the interrelation between the use cases comprising the system. Note that the Standard Products are the objects that can be assigned permanent data locators by the Curation and Reproducibility use case.

4.2 Non-Functional Requirements

In addition to the functional requirements described in the use cases, a set of non-functional requirements is implied by the system design.

4.2.1 Performance

Science ready products will be available in most cases no later than 30 days after the completion of the observation. The system must be sufficiently performant to allow up to three cycles of calibration pipeline processing and one imaging pipeline execution during this period while still permitting time for the quality assurance process.

4.2.2 Scalability

The system must be designed so that hardware may be added and removed to scale the system to match demand. This is true both for the processing hardware and for the infrastructure supporting the user interfaces.

4.2.3 Capacity

The system should be able to support multiple simultaneous users transparently from the users perspective.

4.2.4 Availability

While the system should be available at all times, the SRDP system is not a high availability system. Normal observatory practices should be sufficient to meet the availability requirements.

4.2.5 Reliability

The deployment system shall reliably and efficiently deploy tested code. Code must work in production as it did in testing. The system shall be capable of deploying an identified set of software components and associated configurations, and reproducing that set as needed. The system tracks the version of each component in the set, and maintains a history of component deployments for troubleshooting purposes.

4.2.6 Recoverability

The system shall employ reasonable security precautions to ensure that only authenticated and authorized users may access. The In the case of downtime either planned or unplanned, any processing jobs in the system should be in a known state, and those jobs which were terminated prematurely restarted.

4.2.7 Maintainability

It shall be possible to update or otherwise maintain individual portions of the system without requiring the full system to be redeployed. Similarly to recovery from system outages, any affected processing jobs must be restarted.

4.2.8 Security

The system shall employ reasonable security precautions to ensure that only authenticated and authorized users may access proprietary data, computing resources are accessed only by authorized users, and prevent inadvertent or malicious actions which compromise the system.

4.2.9 Data Integrity

Users must have the capability to ensure that data is not modified, altered, or deleted without authorization in either storage or in transit.

4.2.10 Usability

Users should find the interfaces responsive and intuitive, long actions should produce either busy notifications or status updates. Deep trees requiring multiple clicks to access data should be avoided. Data should never appear "lost" to the users, for example availability of products in the archive search interfaces should be perceived to be instantaneous once the ingest process has completed.

4.2.11 Interoperability

The system shall provide appropriate VO interfaces and be fully interoperable with other VO compliant tools.

4.3 The SRDP Capability Roadmap

The SRDP Project Lifecycle Phases are defined in AD3 with respect to the Initiation Phase, the Deployment Phase, and the Transition Phase. The project phases as defined are intended to establish a release cycle, stabilize the cycle to align with other interdependent releases, and steer all the SRDP processes into normal observatory operations. These phases delineate the project planning, but do not address the Capability Roadmap.

The implementation plan will be continually refined as detailed requirements and implementation are developed, deployed, and feedback is received from stakeholders. Progress will be made on many of the use cases simultaneously, and capabilities deployed to general use as soon as they are thoroughly tested. An initial capability roadmap is presented below to describe the evolution of the project, delineating early- (implementation waves 1 and 2) mid- (waves 2 through 4) and late-project (waves 4 and 5) targets.

Early-Project

The initial objectives of the project are to lay the foundation for more advanced deliverables later in the project, while delivering value to the user community as early as possible. During this period, the standard ALMA pipeline will continue improve as the project seeks to decrease the effort required to perform standard calibration and imaging. The VLA Sky Survey will be observing the second half of the first epoch and delivering the first set of single epoch images.

Nearly all SRDP capabilities rely on pipeline generated science quality calibrations. While these are already available for ALMA, pipeline and process changes are required to support this for the VLA. During the first years of the project, the VLA pipeline will be improved to support science quality calibrations and projects will be processed through the full quality assurance cycle, producing science quality calibrations in the archive.

Ingest and serving of products from large projects can be developed in conjunction with similar requirements from the Sky Survey. This provides a mechanism to leverage the existing science quality products within the community and develop the image and catalog search capabilities within the Archive Interface.

Several longer-term improvements will begin during this period with the SRDP project providing input to a planned redesign of the proposal submission tool (the PST used for the VLA, VLBA, GMVA and GBT). Many of the workflows require capture of additional information at proposal time, and ensuring these requirements are defined and captured as part of the requirements is crucial. Effort will also be dedicated to better defining and characterizing the quality assurance of data products. Although this will likely be an ongoing effort throughout the life of the project and into operations, a significant effort to build a foundation will be allocated in the first few years.

Finally, the Archive Interface and other infrastructure will be developed in these early years. The initial capabilities to be delivered are the Restore use case for both VLA and ALMA, and an initial Optimized Imaging interface for ALMA data. This will make use of the existing ALMA imaging pipeline, and the standard calibrations already provided by the ALMA observatory.

Mid-Project

During the mid-project period, SRDP operations will be increasing with Standard Calibrations for the VLA becoming more automated, and increasing coverage of observing modes and bands. The refactored proposal submission tool will be deployed so the information required for SRDP workflows and template observing will be captured.

The capture of the intents for Time Critical observations in the PST will allow the development of that workflow to continue, a quick look imaging pipeline will be available for some modes of operations, although the automated images probably will not yet be supported.

Interfaces to support the recalibration use case for both telescopes will be developed allowing user specified recalibration. The optimized imaging interface will be expanded to include VLA products as the initial deployment of the VLA imaging pipeline. Introducing optimized imaging first allows the experience gained through this processing to be applied to the development of the standard pipelines. Significant manual and user intervention is expected during this period as the pipeline is improved and refined.

The Curation and Reproducibility use case will be addressed in the mid-project period by the minting of digital object identifiers and creation of landing pages through the Archive Interface.

Late-Project

During the final years of the project, operations will be becoming more routine with standard calibrations commonplace and mature imaging pipelines for both telescopes available. The SRDP operational model will be well established and capabilities for support of Large Projects can be reviewed and extended. The VLA imaging pipeline will be used for the production of standard products, delivering the standard images to the Archive for most project.

Although the development of the image combination pipelines may have begun in the mid-project, availability through the Archive Interface will only come as the project nears completion.