Southern African Large Telescope

Title:	**SALT Data Archiving**
Author(s):	**Alexei Kniazev, Martin Still**
Doc. number:	**2300AC0001**
Version:	**2.0**
Date:	*July 09, 2007*
Keywords:	Archive, Data distribution
Approved:	**David Buckley (Ast Ops Manager)**

Signature:_____ Date:_____

ABSTRACT

In this document the current scheme for SALT archiving is summarized and ideas for possible improvements are discussed. The current version of the document was accepted during SAs+IT meeting.

Contents

1   Introduction

This documents reflects our understanding of data distribution in the file archive.

2   Current Scheme

2.1   Transfer of Observation Data

```
Data flow     scam.salt       pdet.salt    Data flow
    |            |\              /|             |
    |            | \            / |             |
    |            |  \ getfits  /  |             |
    v            |   \        /   |             v
                 |    \      /    |
              rsync qcpc.salt  rsync
                 \   (quick     /
                  \   look)    /
                   \    |     /
                    \ rsync  /
                     \  |   /
                   saltdata.salt
                  (Sutherland storage)
                        |
                  = To Cape Town =
                        |
                      rsync
                        |
                = via internet (VPN) =
                        |
                  ctfileserver.saao
                  (Cape Town storage)
```
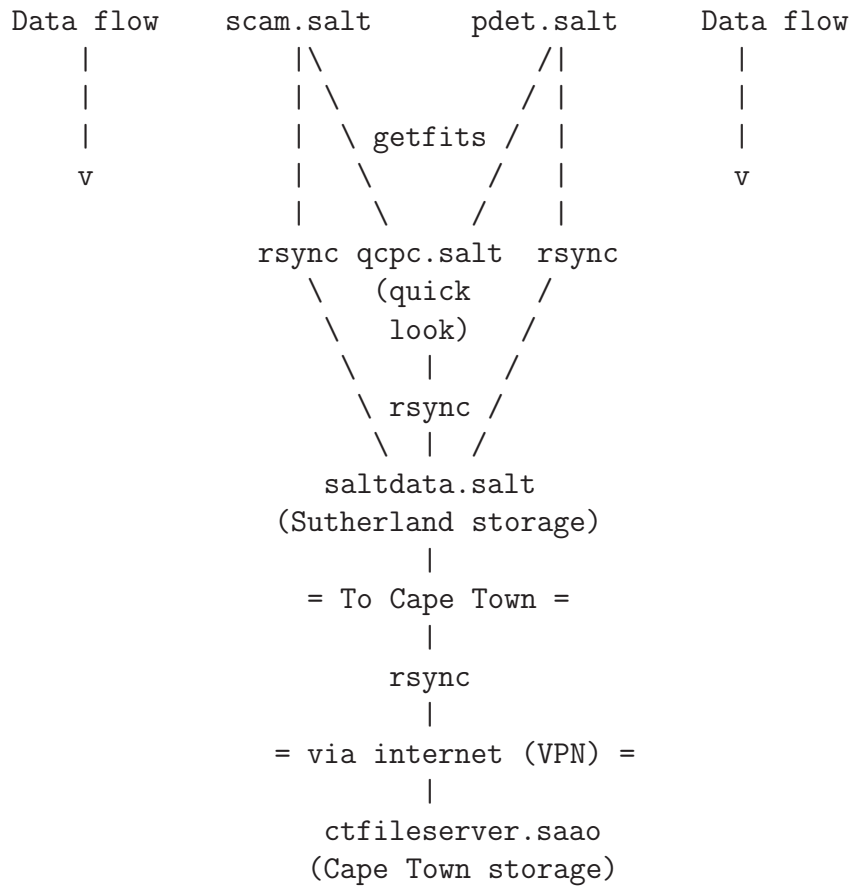
Figure 1: Data Flow from SALT telescope in Sutherland to Cape Town.

The current scheme for copying SALT observational data from the telescope to the Cape Town archive is outlined in the document http://www.saao/it/resources/software-development/projects/system-administration-software-projects/salt-data-download. We summarize the data flow below:

1. Observational data are written to the `scam.salt` and `pdet.salt` data disks at Sutherland directly after each CCD readout. There are directories */data/ccd/data/***obsdate**

on each of these computers, where **obsdate** is a directory created for each night of observations in the format YYYYMMDD and YYYY is a year, MM is a month and DD is a day/night of observations.

2. Data are rsynced to `qcpc.salt` computer for temporary on-fly reduction into directory */data/ccd/data/***obsdate**. Data form both acquisition systems (SALTICAM and RSS) are stored in the same directory.

3. Data are rsynced from all three of the above machines to the `saltdata.salt` machine at the telescope for temporary archiving.

4. Observation data are copied from the `saltdata.salt` computer to the Cape Town computer `ctfileserver.saao` that is mounted on `saltastro.saao`.

From 19:00 - 07:00 data is copied from qcpc.salt to saltdata.salt using a cron job that repeats every 30 minutes. During the same time data is synced down to ctfileserver.saao in Cape Town. At 07:00 every morning data is copied from scam.salt and pdet.salt to saltdata.salt just in case qcpc.salt missed anything during the night. saltdata.salt then sync's data down to Cape Town at 08:00.

2.2   Structure of Raw Observation Data Archive

On the `saltastro.saao` computer data are located in directories
*/salt/***instrum***/data/***obsdate***/raw*
where **instrum** is either `rss` or `scam` depending on the acquisition system. There is no subdirectory structure below this level.

In most cases, raw data is recorded in FITS format. The exception is SLOT mode which is split between a non-standard binary format and a non-standard ASCII file, with name extensions .bin and .head, respectively.

2.3   Structure of Reduced Data Archive

Currently our scheme for archiving reduced data is as follows:

1. On the `saltastro.saao` machine, all reduced data are placed into within */salt/***instrum***/data/***obsdate***/products/* or directories below it.

2. Currently the telescope does not write a complete set of keywords to files. Additional keywords are required to characterize the data and manage the data. Each night, FITS-header corrections are stored in an ASCII file by the duty SALT-astronomer using the format defined in document SALT2350AY0001. The name of each file is `list_newhead_yyyymmdd` and all are stored on the `saltastro.saao` computer in two places: (1) directory */home/sa/Yas_scripts/headchangefiles/* (2) directory */salt/combined/data/***obsdate***/*.

3. All 'raw' FITS-files with corrected FITS-headers are placed in the directories
   */salt/***instrum***/data/***obsdate***/products/rawfixed/*. All FITS-files that do not need
   FITS-headers corrections have symbolic links within this directory.

4. Any SLOT mode observations are converted from binary data to FITS standard and
   are stored in the directory
   */salt/scam/data/***obsdate***/products/rawfixed/slotmode/*;

5. All data resulting from primary pipeline reductions are placed into the directory
   */salt/***instrum***/data/***obsdate***/products/all/*;

6. Reduced SLOT mode observations are stored in directory
   */salt/scam/data/***obsdate***/products/all/slotmode/*;

7. All raw and reduced data are separated by Principal Investigator (PI) name and are
   linked into directories
   */salt/combined/data/***obsdate***/**PI_name**;

8. An ASCII README is also stored in this directory. It contains a list of all program
   files and a summary of the night is given.

9. All raw and reduced SLOT mode data are separated by Principal Investigator (PI)
   name and are linked into directories
   */salt/combined/data/***obsdate***/**PI_name***/slotmode*;

3 Proposed Changes, Additions and Improvements

3.1 Structure of Raw Observation Data Archive

1. Currently the observation data for one night are located in different branches of sudi-
   rectories on `saltastro.saao`:
   */salt/scam/data/***obsdate***/* and
   */salt/rss/data/***obsdate***/*.
   This is not an efficient design because SALTICAM and RSS data can belong to the
   same program, e.g. the slitview image. A proposal to change the directory structure
   has been agreed upon where we have:
   */salt/data/***obsdate***/***instrum***.

   Request to IT: Although note that no action has been taken for many months. The
   SAs request IT to provide an update on this item please. The caveat to this operation is
   that the archive structure for past observations will need to be corrected retrospectively
   to meet the new requirement.

2. **Question to IT:** Do we really need */salt/data/* subdirectory in this scheme?

3. Retrospective access to SLOT-mode data. We agreed finally that database has to refer to both binary and FITS files with SLOT-mode data that have to be stored in the archive, inside of directory with raw data.

   Request to IT: We would like to ask IT to find a solution for this inside of the current archiving scheme.

4. Currently, we cannot make permanent changes to raw FITS keywords on saltastro.saao because the rsync operations will copy the old version of the file from saltdata.salt during the next cron job. Until the telescope is writing all required keywords we need the ability to permanently alter raw files. Currently this is performed by storing some raw files twice on the saltastro.saao machine and this is a relatively poor use of archive space. Do we:

   - Correct FITS-headers once and save changed files in the new directory (current scheme). This increases archive size (negative), but the correct version of the raw FITS-file is instantly available.

   - Since ASCII files containing all of the required corrections already exist, all corrections could be done "on-the-fly". In this way we save disk space (positive), Negative side of this solution is that ANY file could be reached ONLY after a limited time, since the "on-the-fly" correction has to be done first.

   - As outlined below, we cam create meta-tables for a night of data that permanently contain the corrected keywords. This would allow us to delete large redundant FITS files and be able to search the table rather than the raw data, saving both time and disk space.

   Request to IT: The real long-term solution to this problem is to prevent legitimate alteration to raw FITS keywords from being overwritten during rsync operations. Since this needs to be done without risking the permanent deletion of raw data we would like to ask IT to find a solution as **a high priority**.

3.2   Structure of Reduced Observation Data Archive

1. Delete all subdirectories below current directory *products* and put all reduced data into two directories. Reduced files have to be located in:
   */salt/data/***obsdate**/**instrum**/*product*

2. Create a FITS table in each of:
   */salt/data/***obsdate**/**instrum**/*product*
   */salt/data/***obsdate**/**instrum**/*raw*
   The table will contain all of the information required to ascertain the contents of each FITS file and which PI or program it belongs to. As a first cut this table could contain all keywords but could potentially be trimmed down in the future. Benefits of these files are relavent to the SAs, the pipeline, archive searching and PIs:

- Provide a comprehensive reference for file content without having to open large files to understand their content.

- Provide the basis for observing statistics.

- Provide an easy resource for comparing archive data to the expectations of the database.

- Speed up the pipeline by providing an efficient means of data identification.

- In conjunction with a filtering tool within the data reduction/pipeline software package, allow the user to filter on any keyword, create file lists and create sub-tables in different formats e.g. ASCII.

- Sub-tables can be prepared and distributed to each PI. Each PI can use the table and tool to create e.g. a list of biases, or a list of all V images between 01:00 and 01:30 in FT mode.

- These tables will be created after raw keywords have been fixed, so a permanent and easily accesible record of the correct keywords is always at hand even if the corrected files themselves are over-written by the cron jobs.

3. We decided to add to the current READMEs in ASCII-format also pipeline-generated documentation at e.g.
http://saltastro.saao/PySALT/docExample/PipelineLog.html.
Both of them have to be sent to PIs.