File Naming Convention for the NOAO Science Archive

R. Seaman, E. Stobie, F. Economou, F. Valdes, D. Herrera
NOAO Science Data Management

Overview

NOAO data sets have historically been renamed with a unique serial number identifier before archive ingest to prevent the submission of duplicate file names to the archive. Duplicate file names for the original raw data are always a possibility. For example, most instruments have a test mode that reuses the same file name over-and-over, and almost all instruments permit observers to reuse observation names in one way or another. However, when each individual file is later given its own unique serial number identifier it is impossible to associate multiple files that derive from a single observation without reading the headers (or other metadata) of the files.

The Hubble telescope defined a simple pattern for determining file names known as IPPPSSOOT, where I represents the instrument, PPP the program ID, SS and OO the observation set ID and the observation ID, respectively. Finally, T represents the source of transmission. Appended to the ipppssoot is the data type qualifier: _raw, _cal, _msk, etc. The HST naming problem is a bit more straightforward than NOAO's since there is only one observatory/telescope, and with a very highly structured observing system the chance of duplicate files is nil.

However, NOAO can define an identifier using similar principles to HST's to make each file name unique while sharing the majority of the file root among several associated files. The intent is to ensure that each name is indeed unique while in just a few characters giving as much information about the content of the file as possible. The goal is an identifier that is determined from one or more keywords in the file header so that according to a set of rules the file name can be heuristically derived from these header keywords and will never change. For most instruments a candidate unique identifier is an observation ID based on the date and time of the observation. The filenames begin with the components: observatory, telescope, and instrument (see Table 1). This is followed by the UTC date and time, and single character observation type, processing and product types (see Tables 2-4), perhaps a short filter ID, and the file extension, typically ".fits.fz".

Rather than refer to this scheme as something unpronounceable like STIDTOPP, this document will simply call it the NOAO Data Set Identifier or DSID for short.

Table 1 – Site / Telescope / Instrument prefixes: Each raw and pipeline reduced file name will have a prefix conveying the observatory site, telescope, and instrument (or class of instrument). Grayed-out instruments have not been scheduled since the new names were deployed. These and other configurations (and at least three more telescopes) are present in archive holdings.

Site	Telescope	Instrument	Type	Prefix
Cerro Pachon	SOAR	Goodman	spectrograph	psg
		OSIRIS	IR imager/spectrograph	pso
		SOI	imager	psi
		Spartan	IR imager	pss
		SAM	imager	psa
Cerro Tololo	Blanco 4m	DECam	imager	c4d
		COSMOS	spectrograph	c4c
		ISPI	IR imager	c4i
		Arcon	imagers/spectrographs	c4a
		Mosaic	imager	c4m
		NEWFIRM	IR imager	c4n
	1.5m	Chiron	spectrograph	c15e
		Arcon	spectrograph	c15s
	1.3m	ANDICAM	O/IR imager	c13a
	1.0m	Y4KCam	imager	c1i
	0.9m	Arcon	imager	c09i
	Lab	COSMOS	Spectrograph	clc
Kitt Peak	Mayall 4m	Mosaic	imager	k4m
		NEWFIRM	IR imager	k4n
		KOSMOS	spectrograph	k4k
		ICE	Opt. imagers/spectro.	k4i
		Wildfire	IR imagers/spectro.	k4w
		Flamingos	IR imager/spectrograph	k4f
	WIYN 3.5m	WHIRC	IR imager	kww
		Bench	spectrographs	kwb
		MiniMo/ICE	imager	kwi
		(p)ODI	Imager	kwo
	2.1m	MOP/ICE	imager/spectrograph	k21i
		Wildfire	IR imagers/spectro.	k21w
		Flamingos	IR imager/spectrograph	k21f
		GTCam	imager	k21g
	Coude Feed	MOP/ICE	spectrograph	kcfs
	WIYN 0.9m	HDI	imager	k09h
		Mosaic	imager	k09m
		ICE	imager	k09i
Any	All	Other	default	alt

Implementation Notes

The requirements and boundary constraints are that the new DSID should:

1. be guaranteed unique,
2. be as brief as practical,
3. be predictable from controlled metadata,
4. and have a root name shared between different data products associated with the same observation.

The resulting identifier looks like:

- *<obs_tel_instr>_<UTC_date_time>_<obstype><proctype><prodtype>.<ext>*

Examples for raw data are:

- psg_131212_042025_ori.fits.fz
- k09h_131212_061232_zri.fits.fz
- c15e_131212_092837_cri.fits.fz
- k4n_131212_102508_dri.fits.fz
- c4d_131212_192824_fri.fits.fz

Each of these files is a tile-compressed FITS file. Since UTC dates and timestamps are used, the data from multiple telescopes can be assembled into a single time sequence as here. The 3 or 4 character prefix describes the site ("p" for Cerro Pachon and "c" for Cerro Tololo in Chile, "k" for Kitt Peak in Arizona), the telescope (SOAR, the WIYN 0.9m, the SMARTS 1.5m, Mayall and Blanco), and the instrument (the Goodman spectrograph, HDI, the Chiron echelle, NEWFIRM, and DECam). And the 3 character suffix for each describes the observation type (object, zero, comparison, dark, flat), as well as that each is a raw image ("ri"). The various codes are listed in Tables 1-4.

For pipeline-reduced data, the idea is to build new names off the file root for the original object exposures, for example:

1. k4n_131212_090458_ori.fits.fz
2. k4n_131212_090458_odi.fits.fz
3. k4n_131212_090458_osi.fits.fz
4. k4n_131212_090458_odg.png
5. k4n_131212_090458_ose.fits.fz

The first file listed is the raw object exposure. Number 2 is the pipeline-reduced and resampled image and the third is a stacked image from an exposure sequence. Number 4 is a PNG graphics image of the resampled image, for example a preview. And number 5 is an exposure map for the stacked image.

Each of the fields of the DSID will be discussed separately below to provide context for generating new codes and new identifier formats to serve different purposes (for example, for survey-reduced data products). Most of these fields are already supported by the iSTB DT/SB keywords or by the NHPPS set of keywords. A key requirement will be to provide a useful default value (either explicit or implicit) that will apply when the corresponding metadata for a particular field is not defined (or is simply missing) for a particular data product.

Observatory / Telescope / Instrument token

This token will be a mapping to the current DTOBSERV, DTTELESC and DTINSTRU keywords that are added to raw data by iSTB, see Table 1. The length of the token is required to be as brief as possible without sacrificing human readability. The telescope and instrument identification is known from the client data acquisition host.

In some cases the instrument is ambiguous because several configurations use the same data acquisition computer. The recently deployed disambiguation of the Spartan and Goodman instruments at SOAR, which are served from the same host, should provide a good foundation for other such instances. In most such cases, however, it is less critical to distinguish the instruments and it may be acceptable to use a common default value.

The telescope is always well known (except perhaps during commissioning).

The observatory is an institutional question, and is thus subject to change. For instance, operations of the Kitt Peak 0.9m telescope shifted from KPNO to the WIYN partners. The telescope site is of more permanence and is of scientific importance even in rare instances of a telescope being moved (*e.g.,* the Burrell Schmidt). Thus a letter denoting the site will be used rather than the observatory itself.

UTC Date / Time stamp

The implementation derives this value from the DTUTC keyword, which was added as a result of a reliable timekeeping initiative several years ago. DTUTC is precise at the 1s level. Alternate arrangements will be needed if any of our current or planned instrumentation is capable of more rapid exposures. In the past, certain IR instruments have at least theoretically been capable of observing cadences faster than one second and the solution at that point was to add a disambiguating serial number controlled by the instrument software.

There is no reason that a couple of digits of precision couldn't be added to DTUTC, though this may be a concern on some legacy data acquisition computers. In that case the notion might be to add a couple of zero digits.

For the sake of brevity we will omit the two century digits. If 20th century or post-21st century data products ever need be supported, the obvious clarification will be that in the absence of the century digits that 21st century data sets are implied.

An alternative proposal would be to use DTCALDAT instead of DTUTC since NOAO archive data are batched by the observing calendar date not by UTC. In that case, the timestamp might be replaced by the SB_RECNO. In such a case the ordering of the fields might differ. This might be considered in the future for certain instruments or survey data sets.

Observation type

The observation type is specified for some instruments but not for others. The keyword used varies, with instances of IMAGETYP, OBSTYPE, IMGTYPE and OBS_TYPE (and none). An added complexity is that some instruments, particularly in the IR, use science data to construct calibration frames through median filtering, etc. A flexible set of default(s) will be needed to cover the edge cases.

Table 2 – code letters for observation types

code	Observation type
o	Object
p	Photometric standard
z	Bias
f	Dome or projector flat
s	Sky flat
d	Dark
c	Calibration or comparison
i	Illumination calibration
g	Focus
h	Fringe
r	Pupil
u	Unrecognized

Processing type and product type

For pipeline data sets these should be well specified. For survey data sets care must be taken to ensure the same. For raw data the PROCTYPE is just a fixed value, "r" for raw, while the PRODTYPE will be "i" for image. A default, "u", will support cases that are either unspecified or unknown.

Table 3 – code letters for processing type

code	Processing type
r	Raw
o	InstCal
c	MasterCal
p	Projected
s	Stacked
k	SkySub
u	none of the above

The list of product type is likely to grow in the future, and some types such as "g" for graphics may split into subclasses, for instance, "p" for preview versus "t" for thumbnails. An alternate concept is to append a digit (for "g" only) to distinguish smaller or larger previews.

Table 4 – code letters for product type

code	Product type
i	Image
j	Image 2nd version [1]
d	Dqmask
e	Expmap
g*N*	Graphics (*size*)
w	Weight
u	none of the above

File extensions

These currently include "fits", "fits.fz", "fits.gz", "png", "xml", and "hdr", and the list can be expected to grow. The idea of the DSID is that the same root observation ID will be appended with different proctypes and prodtypes (and thus extensions) to match. Not all roots will correspond to the existence of all file extension types. For example, a raw exposure will generally have only a tile-compressed FITS file and the corresponding text HDR file. Whereas a pipeline-processed

[1] Product type 'j' denotes a second version of an image, e.g., for DECam this is a second version of a stack with more aggressive sky subtraction.

(nonstacked) image will have a science exposure, different kinds of masks, and different kinds of PNGs (preview and thumbnail).

Delimiters

Field delimiters in identifiers are underscores, except for the file extensions that use the usual unix dots. The OBSTYPE, PROCTYPE and PRODTYPE are a joint field omitting the delimiter. It may be desirable to construct other fields to also be self-delimiting such that for brevity the identifier could have a short form omitting the delimiters entirely.

Alternately the delimiter between the UTC date and time (an atomic combination) could be the ISO 8601 standard "T", but readability and simplicity argue for a single delimiter throughout.

Optional fields

The remaining fields are optional. Each optional field will be prepended with an explicit underscore delimiter, thus permitting multiple character tokens to be used. It is anticipated that any optional fields will always be added in a standard order, but it must be possible to reliably identify each field from its unique format.

A complete DSID looks like:

- *<Obs_Tel_Inst>_yymmdd_hhmmss_<type_codes>[optional_fields].<exts>*

The optional fields are currently envisioned to include at least three types of information: filters, sub-images, and versions.

Filters

The filter field presents two challenges. First, some instruments do not capture this metadata. Second, other instruments have a very rich filter set, especially in narrow bands. For the latter situation NHPPS has already addressed this question and we should seek to benefit from prior art. For the former issue the only practical option may be to choose a default value indicating "unknown", although it may be possible to indicate a filter wheel position, rather than the filter ID itself. An unknown filter should perhaps be separate from defaults indicating "none" or "not applicable". After the initial underscore, a filter field can be any NOT purely numeric field, with or without a "v" prefix, but without embedded whitespace. Filter identifiers are the only part of the identifier that is case sensitive. Example: "_V" is ok, but "_v2" would be ambiguous with a version specification.

Sub-images

Some instruments can create multiple images per exposure. Others can take data more rapidly than one image per second. Thus some mechanism is needed to distinguish between these multiples. A sub-image field will be purely numeric (after the initial underscore). If both are present, the sub-image will follow the filter. Example: "_12" would mean the twelfth sub-image. Whether sub-images are zero or one-indexed is left as a case-by-case determination. For instruments that create multiple images per exposure, a DSID without a sub-image field will refer to the entire set of per-exposure sub-images as a whole.

Versions

Pipeline and survey data sets will occasionally be reprocessed and there must be a way to reflect such instances. An identifier's version field will begin with "_v" (underscore-lowercase v) and will be purely numeric after. The version field will follow both filter and sub-image, if they are present. Versions will be explicit. If an initial version is later supplanted by new data files, the original identifiers will continue to correspond to the original data files.

Usage

It is important that the NOAO data set identifiers be applicable to all NOAO Science Archive holdings, including survey data in addition to raw and reduced observations. The survey case would include other data products that may arise in the future. Not only may there be holdings that require the list of Obs_Tel_Inst combinations be expanded, but files may require a different composite namespace entirely, *e.g.,* for a digital representation of a published catalog compiled from many sources and on many dates. In this case the remaining fields may also not apply. In that case, a separate prefix token (as opposed to <Obs_Tel_Inst>) can trigger a completely different naming scheme than the baseline DSID.

Constructing raw DSIDs

Building a DSID from scratch requires several pieces of information:

Table 5 – information required to create DSID – fields in *italics* are optional

Information	Source
Site	DTSITE
Telescope	DTTELESC
Instrument	DTINSTRU
UTC date / timestamp	DTUTC
Type of observation	OBSTYPE, IMAGETYP or IMGTYPE
Type of processing	PROCTYPE
Type of data product	PRODTYPE
File format	Created by iSTB or pipeline
Compression	Applied by instrument or iSTB
Filter ID	*Known by pipeline*
Sub-image number	*Inferred from instrument*
Version ID	*Known by pipeline*

For raw data the site, telescope, and instrument are inferred by iSTB from knowledge of the data acquisition host, and in practice various configurations in Table 1 are degenerate against Table 5. A standalone tool would likely be configurable by command line switches to infer these fields in different combinations.

The DTUTC post-exposure timestamp has been present in the NOAO Science Archive for several years. For holdings prior to this a reliable date/timestamp is inferred from different keywords, e.g., DATE-OBS, UT, DATE, etc., or perhaps even from the unix file creation timestamp. The precedence is DTUTC then DATE-OBS. If the instrument provides DATE-OBS in the full date/time format then this is sufficient for both the date and time, but otherwise the time inferred from either the UT keyword, the UTC keyword or TIME-OBS (in that order). If

DATE-OBS is in the pre-Y2K format it will also be interpreted correctly, but for at least one instrument (ISPI at the Blanco c. 2004) the DATE-OBS format is incorrect both pre and post Y2K (MM/DD/YYYY). The DSID parser figures this out. Finally the DATE keyword is consulted. The date and the time can separately default to values of "NODATE" and "NOTIME".

Different instruments use different keywords to convey the type of observation. The heuristic implemented by iSTB is to accept values from the OBSTYPE, IMAGETYP, or IMGTYPE keywords – and those values comprise several aliases for each type of observation. These keyword variations and these aliases were compiled from several hundred thousand images over the past two years or so. Even so, it is quite likely that legacy holdings contain additional variations of keywords and of their values that will need to be integrated in the future.

The processing type and product type default to "ri" for all raw data. File format and compression are unchanged using DSIDs versus iSTB serial numbers. Filter IDs, sub-images, and versions are currently unused for raw data.

Deriving pipeline DSIDs

While the work of creating raw data DSIDs rests with the top half of Table 5 (through the observation type), the pipelines are responsible for the bottom half of the table, in particular for introducing knowledge of the processing and product types and of the resulting file formats. Pipelines also often result in reprocessing data sets and thus create the possibility of introducing a new version number – and finally, since a pipeline needs to differentiate between filters it provides an opportunity to capture that information as a brief token suitable for embedding in the resulting DSIDs.

A pipeline is thus responsible for:

1. retrieving the DSID from the DT_RTNAM (*e.g.,* k4k_131213_074155_ori),
2. stripping the "ri" from the DSID,
3. appending the appropriate code letters from Tables 3 and 4,
4. possibly appending an optional filter and/or version field,
5. appending the appropriate file format extension, and
6. writing this value to a new keyword, PLDSID.

Passing pipeline and survey DSID to iSTB

When a FITS file containing a PLDSID keyword is submitted to iSTB by a pipeline or in a batch from a survey, iSTB will create an output file with that name in the same directory as usual (that is, in the directory …/PLQUEUE/PLQNAME for a pipeline, or in SURVEYID/SURVEYD1/SURVEYD2). Compression and other handling such as extracting files from FITS foreign encapsulation will proceed as always. If there is a name collision for whatever reason, the file name will default

to the usual unique iSTB serial number. In this case the offending PLDSID keyword will remain in the image header to aid in resolving the issue after the fact.

Utility scripts for generating DSIDs

A perennial issue is that observers specify their own observation file names for almost all NOAO and partner instrumentation, and that they later wish to rename data holdings from the NOAO Science Archive to the original names. This requirement will not change with DSIDs as opposed to the legacy serial numbered files. SDM currently provides scripts so that user can perform this chore for themselves.

Conversely it will also be necessary to restore DSID names after the fact to files that have become renamed, whether to the original names or others. This will be used, for instance, to batch-rename the large backlog of serial numbered files to the DSID convention. A script or compiled command to implement DSID renaming will require more detailed knowledge of NOAO's telescopes and instrumentation, and of the variations of observation, processing and product types among these.