

# Essential Imaging Metadata

---

Richard A. Shaw (DPP, NOAO/Tucson)

8 May 2007

*Metadata for image data products produced by DPP are considered science-critical if they are essential for a correct and complete scientific interpretation of the data. Such metadata can be usefully categorized in terms of “coverage” in a Virtual Observatory sense. This document establishes which metadata are considered critical, and therefore must meet the highest standard of reliability before the data are made available for general use. Some validation techniques and remediation heuristics are also suggested. The status of validation for critical metadata must be tracked in the DPP E2E system. Finally, a set of one or more configured reference tables should be constructed to support metadata validation within the E2E system.*

## 1 Introduction

Metadata, or data that describe data, are collected or created at various points in the planning, acquisition, and processing of science data that are obtained through NOAO observing programs. Much of the metadata that are available to the user are stored in the headers of science data (FITS) files in ASCII form, as keyword-value pairs. Metadata fall into a variety of categories, but perhaps the most relevant from a user point of view are the following:

- **Telescope/Environment:** including telescope name, configuration information, observer name and proposal identifier, celestial coordinates of the telescope, airmass/zenith distance, tracking information, ADC use, etc.;
- **Instrument/Detector:** metadata that are required for generating Level-2 products from Level-1 raw data, and include the instrument configuration, detector/amplifier designations, gain settings and anticipated read noise, operating temperatures, portion of the image area devoted to astrophysical targets, filter/disperser or other optical elements, pixel scale, etc.;
- **Science/Calibration:** including the World Coordinate System mapping, field-of-view, resolution or delivered image quality, start of the exposure and its duration, effective bandpass and sensitivity of the system, brightness units and dynamic range, background level, and transformation to a standard photometric system;
- **Pedigree/Provenance/Quality Assurance:** including version of the control software, version of data processing/calibration software, bias stability, processing steps performed, calibration reference files used, etc.

These metadata serve a variety of purposes, from recording status information, calibration dependencies and processing steps, or the results of calibrations. In practice only a modest subset of the metadata are essential to understand a data product from a scientific point of view, particularly for calibrated data, and that is the focus of this report. Other metadata may be important or even critical to the proper functioning of the DPP end-to-end system, but most such metadata are not considered here unless they impact the scientific use of the data.

## Data Products Program

Metadata for data products produced by DPP are considered science-critical if they are essential for a correct and complete scientific interpretation of the data. This document establishes which metadata are critical for imaging data, and specifies the metadata, expressed as science image header keywords, that must meet the highest standard of reliability. The science-critical metadata can be usefully categorized in terms of “coverage” in a Virtual Observatory sense—i.e., the extent to which the data product spans each of four fundamental dimensions for a dataset: space, time, wavelength or energy, and dynamic range. (Another dimension, polarization, is not considered here because NOAO currently has no polarimetric imagers.) If these critical metadata are missing or incomplete it will not be possible to match NOAO datasets in one or more of these dimensions in response to VO queries. Metadata that are present but with incorrect values obfuscate the scientific use of the data, confuse and frustrate the intended users, and in the worst case render the data scientifically useless. From the programmatic point of view, even if some metadata are missing, incomplete, or inaccurate, the confidence in all of the NSA holdings is greatly eroded. Thus, assuring the completeness and accuracy of the science-critical metadata is of the utmost urgency if datasets provided by the NOAO Science Archive are to be valued by the scientific community.

The following sections describe the relevant metadata for each dimension of coverage, and include a table of keywords and definitions. Note that many of these keywords and their definitions are specified in form, format, and content by the FITS standard (IAUFWG 2005). Valdes (2000) provided a detailed specification of keywords that are used for the NOAO Mosaic cameras. Each table below also establishes a level of criticality to each keyword, defined as 1=critical, 2=important, 3=informational.

- Keywords with a criticality of “1” must be present in all headers (i.e., for Level-1 and higher data products), and values must be validated and corrected if necessary.
- Keywords with a criticality of “2” must be present in all headers (i.e., for Level-1 and higher data products), but values must be validated and corrected only required if Level-2 and higher products are made available to general VO users.
- Keywords with a criticality of “3” should be present in the image header, and must be correct if present.

Regrettably, metadata that were written to legacy NOAO datasets are known to occasionally be missing, incomplete, or inaccurate. In most cases it is possible to validate the metadata automatically, and in some cases it is possible to correct the metadata if an error is detected. Each section also includes rules that should be applied to validate critical metadata, and strategies that can be followed to correct faulty metadata.

## 2 Temporal Coverage

The interval of time over which the detector was exposed to the astrophysical source defines the temporal coverage of the data. Generally this is characterized in the metadata as the exposure start time and the duration of the exposure. In principle one must also record (or correct for) effects such as “dead time” intervals, e.g., when the shutter is closed but the detector is still able to record events such as cosmic rays, and exposures that are interrupted or paused for some reason during the course of an exposure, but are later completed.

The following table lists the header keywords that are related to temporal coverage, their level of importance, their definition, and guidelines for validating the keyword values.

**Table 1: Keywords for Temporal Coverage**

Keyword	Crit.	Definition	Validation
TIMESYS	1	The principal time system for all time-related keywords.	Usually this is UTC, but in any case it must be one of the official time systems defined in Appendix D of the FITS standard.
DATE-OBS	1	Date & time of observation start. Format is defined in Appendix D of the FITS standard as: 'yyyy-mm-ddThh:mm:ss.sss'	Valid ranges: $1993 \leq \text{yyyy} \leq$ current year, $1 \leq \text{mm} \leq 12$ , $1 \leq \text{dd} \leq 28$ to 31; see TIME-OBS.
TIME-OBS	2	Time of observation start, which is the part of the DATE-OBS string following the "T".	Valid range: $0:00:00.000 \leq t < 24:00:00.000$ .
MJD-OBS	2	Time of observation start in the Modified Julian Day time system. Note that the precision must be finer than 1 s, but in any case must be sufficient to reflect the accuracy of the data-taking system.	Must exceed 48987 <sup>a</sup> ; derived from DATE-OBS.
EXPTIME	1	Effective exposure time, in seconds.	Must be non-negative. Exposures in excess of 10,000 s should be viewed with suspicion.

Note that it may be impossible to recover from a corrupted or missing date or time, except by examining an observing log (if available) or by comparing with the scheduling database and with the creation dates of other science data files in the observing run. Corrupted date or time metadata makes remediating positional metadata much more difficult or impossible.

### 3 Spatial Coverage

The spatial coverage for images is defined by the World Coordinate System (WCS), the FITS standard for which is described in detail by Calabretta & Greisen (2002). The WCS defines a mapping between the detector pixels and physical coordinates on the sky. Generally speaking, the WCS specifies the coordinate system that is used, the mapping function (i.e., the projection of the sky onto pixels), and a set of keywords whose values contain the coefficients for the named mapping function. There are usually a number of other keywords present in science image headers that, for convenience, provide related quantities such as telescope position, physical pixel size, footprint of the field of view, etc., but these quantities are derived from the WCS and do not fully define the spatial coverage.

The following table lists the header keywords that are related to spatial coverage, along with their definitions and criticality. Note that some of the keywords refer to one of two image axes, which are denoted in the table with indexes *i* and *j*. The convention is that the first image axis is the one for which the index in the data array changes most rapidly.

**Table 2: Keywords for Spatial Coverage**

Keyword	Crit.	Definition	Validation
---------	-------	------------	------------

<sup>a</sup> MJD for the beginning of 1993, the year that electronic storage of data via STB began.

## Data Products Program

Keyword	Crit.	Definition	Validation
NAXIS <i>i</i>	1	The number of pixels along the image axis <i>i</i> .	Must be non-negative integer.
RADESYS	1	Name of the reference system in which the world coordinates are expressed.	Usually FK5, but must be one of the reference system IDs given in Calabretta & Greisen (2002).
EQUINOX	1	Equinox in years for the celestial coordinate system in which the positions are expressed.	Most likely 2000., but should be a decimal value $\geq 1950$ .
CTYPE <i>i</i>	1	Name of the coordinate represented by axis <i>i</i> .	Often RA---TNX or DEC--TNX; must either use the TNX convention or one of the coordinate types given in Calabretta & Greisen (2002). Each sky coordinate must be represented once and only once in any extension header.
CRPIX <i>i</i>	2	Location of the reference point along axis <i>i</i> in units of pixels.	Floating-point number. The reference point need not lie within the physical image array.
CRVAL <i>i</i>	2	Value of the world coordinate at the reference point for axis <i>i</i> , in degrees.	Valid range: $0. \leq \text{CRVAL} < 360$ . For RA; $-90. \leq \text{DEC} \leq 90$ . for Dec.
CD <i>i_j</i>	2	Transformation matrix from pixel coordinates to intermediate world coordinates; CD <i>i_i</i> is the pixel scale for axis <i>i</i> .	$0.01 \leq  \text{CD}_i_i  \leq 100.0$ . Should be verified with a look-up table, indexed by instrument, telescope, and focal station.
RA	2	Right ascension of the center of the instrument FOV in hours (sexagesimal format).	Valid range: $0:00:00.000 \leq \text{RA} < 24:00:00.000$ .
DEC	2	Declination of the center of the instrument FOV in degrees (sexagesimal format).	Valid range: $-90:00:00.00 \leq \text{DEC} \leq 90:00:00.00$ .
WAT <i>i_nnn</i>	3	Description of the non-linear portion of the transformation from detector (pixel) coordinates to world coordinates for axis <i>i</i> . This character string contains coefficients for a polynomial; the length of the string is such that it must continue for <i>nnn</i> FITS header records.	Should be verified with a look-up table, indexed by instrument, telescope, and focal station.
TELRA	1	Right ascension of the telescope position on the sky in hours (sexagesimal format).	Same as RA.
TELDEC	1	Declination of the telescope position on the sky in degrees (sexagesimal format).	Same as DEC.
PIXSCAL <i>i</i>	3	Pixel scale along axis <i>i</i> , in arcseconds/pixel.	$0.001 \leq \text{PIXSCAL}_i \leq 100$ . Should be verified with a look-up table, indexed by instrument, telescope, and focal station.

Note that recovering from missing or incorrect WCS metadata may be impossible if the TELRA or TELDEC keyword values are missing or corrupted. In these cases it may be helpful to consult observing logs or examine the value of the OBJECT keyword for clues of the approximate telescope pointing. It is possible to construct other consistency checks for the pointing—e.g., the zenith distance can be verified from the telescope RA & DEC: if it exceeds 85° it would fail validation; exceeding a sensible value, 75° say, would be cause for suspicion. The best and surest way of validating WCS data is to attempt a WCS solution; *we recommend that this operation be performed on all raw images from CCD cameras if no calibration pipeline is available for this purpose.*

## 4 Spectral Coverage & Photometric Transformation

The interval in energy (or equivalently, the wavelength or frequency interval of detected photons) over which the system is sensitive defines the spectral coverage. This information is important both for cross-matching with other datasets, and for deriving a transformation between the instrumental response and one or more standard photometric systems. The following table lists the header keywords that are related to spectral coverage and photometric transformations.

**Table 3: Keywords for Spectral Coverage & Photometry**

Keyword	Crit.	Definition	Validation
FILTER	1	Designation (i.e., the serial number) for the filter used for the observation. May also include the filter name in the string.	Must be one of the filter designations from the observatory used to obtain the data.
FILTNAME	1	Filter name.	Must be one of the filter names from the observatory used to obtain the data.
PHOTCLAM	2	Average wavelength ( $\text{\AA}$ ) of the bandpass, $\lambda_c = \frac{\int \lambda P(\lambda) d\lambda}{\int P(\lambda) d\lambda}$	Must exceed 3200; derived from look-up table indexed by filter name and instrument.
PHOTBW	2	RMS width ( $\text{\AA}$ ) of the bandpass, $\text{Width}_{RMS} = \sqrt{\frac{\int (\lambda - \lambda_0)^2 P(\lambda) d\lambda}{\int P(\lambda) d\lambda}}$	Derived from a look-up table indexed by filter name and instrument.
PHOTFWHM	2	Width of the filter bandpass ( $\text{\AA}$ ), determined from the difference in wavelength between the points where the transmission reaches 50% of the peak value.	Derived from a look-up table indexed by filter name and instrument.
PHOTFLAM	3	Inverse sensitivity ( $\text{erg.cm}^{-2}.\text{s}^{-1}.\text{\AA}^{-1}.\text{count}^{-1}$ ).	Non-negative.
PHOTFILT	2	Filter used to define the PHOTDPTH and MAGZERO keywords.	Derived from a look-up table indexed by filter name and instrument.
PHOTCOEF	3	Coefficients for transforming from the reference photometric catalog to the photometric system identified by the PHOTFILT keyword.	
PHOTDPTH	2	Brightness of a point source corresponding to a 5-sigma detection above the sky background, expressed in magnitudes.	Non-negative, < 40.
MAGZERO	2	Brightness of an object in the observed bandpass that would produce one count in the science image, expressed in magnitudes.	Non-negative, < 40.

Filter identification is extremely difficult to validate *a posteriori*, although the observing logs may be helpful for this purpose. If data from the same field were obtained on the same night and with a different

## Data Products Program

filter, it may be possible to compare colors of field stars against published catalogs. Note that presence of the new keywords PHOTCLAM and PHOTBW, which replace the tentative use of LAMSTART and LAMEND keywords now produced by the Mosaic reduction pipeline, and the addition of the PHOTFLAM keyword. The intent is to provide a clearer definition of the bandpass that is also compatible with synthetic photometry software (e.g., **synphot**).

## 5 Brightness Coverage

At present there is only a draft IVOA standard for expressing the brightness range covered by an observation, although it would be useful to have this information when searching for an overlapping range in brightness among datasets. The only metadata that is available currently is contained in the PHOTDEPTH keyword, which gives an estimate of the faint end of the brightness range. The upper end of the brightness range could easily be estimated from the detector dynamic range (i.e., the maximum number of counts where the response is still linear) and a characterization of the delivered image quality for the exposure. The definition of this quantity would be the brightest point-source in the observed bandpass that could be measured without exceeding the detector linearity, expressed in magnitudes.

## 6 Other Essential Metadata

Some other metadata are required to describe the origin and pedigree of the data products. While these metadata are less critical to the scientific interpretation of the data, some are still important for users who wish to calibrate raw data from scratch.

**Table 4: Keywords for Other Essential Metadata**

Keyword	Crit.	Definition	Validation
OBSERVAT	1	Name of the observatory where the data were obtained.	For data originating from NOAO systems the value should agree with the location of the machine that provides the raw data.
TELESCOP	1	Name of telescope where the data originated.	Same as for OBSERVAT .
INSTRUME	1	Name of the instrument that generated the raw data.	May be taken from DTINSTRU, which is currently added by iSTB.
OBSTYPE	1	Type of observation, which is either a science target or some type of calibration.	Must be one of: object, bias, dark, dome flat, twilight
PRODCYTP	1	Type of data product.	Must be one of those defined for DPP; an initial list is given in Table 5.
DS_IDENT	2	String to be used in journal citations to uniquely identify this dataset.	See Shaw & Miller (2007).
VO_IDENT	2	Unique URI of this dataset for Virtual Observatory Access.	See Shaw & Miller (2007)
BUNIT	3	Brightness units of values in data raster.	Generally ADU for raw data, and photon/(s.pixel) for reduced; in any case the units string must follow the convention in Greisen & Calabretta (2002).
DQIWIDTH	3	Full-width at half-maximum, in pixels, of point-source spatial profile in the image.	[Measured from the image.]

Note that the OBSTYPE keyword is particularly important to validate. The NOAO data rights policy

specifies that all internal calibration data, such as darks, flat-fields, and arc-lamps will not be considered proprietary, and will be made available to the public immediately upon ingest into the NOAO Archive. In order to enforce the correct proprietary periods, the `OBSTYPE` must be validated. It is easy to identify “bias” (or “zero”) exposures in that the value of the `EXPTIME` keyword will always be zero. Exposures of type “dark” will have finite exposure times, but the images will have uniformly low values. Twilight or dome-flat exposures will have uniformly high values, and no science targets. It may be necessary to visually inspect the flat-field exposures (or thumb-nails) to distinguish them from science exposures.

A set of defined data product types will be established for DPP, and documented elsewhere. The initial list of products is given in Table 5, with abbreviated descriptions. More detailed descriptions of the products and how they are used for each instrument can be found in the NOAO Data Handbook (Shaw 2007).

**Table 5: Data Product Types**

Product Type	Level	Definition
Raw	1	Raw data frame as obtained at the telescope, with additional keywords added
CalFrame	2	Single-frame calibrated
FrameMask	2	Bad pixel mask for CalFrame products
CalReproj	2	Calibrated, re-projected, tiled frame
ReprojMask	2	Bad pixel mask for CalReprj products
Bias	2	Bias structure image
Dark	2	Dark current response template image
DomeFlat	2	Dome flat-field template image
TwFlat	2	Twilight flat-field template image
SkyFlat	2	Sky delta-flat-field image
PupilTempl	2	Pupil ghost template image
FringeTempl	2	Fringe pattern template image

## 7 Validation Tracking

It is essential to track whether the critical metadata have been successfully validated. This could be done at the individual keyword level, but at a minimum the categories of scientific metadata (time, space, spectral coverage, photometric transformation, brightness) need to be tracked. Tracking need not be reflected in the headers of data files, but could be tracked in a database. Values for the fields should indicate the status of the validation: *pass*, *fail*, or *unknown*.

Much of the metadata can be validated, or the range of valid values can be restricted, using simple look-up tables. For example, the pixel scale at a given telescope focal station is well known for all imaging instruments, and the list of valid filters (and their bandpass specifications) is known for each observatory. It would be straightforward for validation software to make use of such tables. The

## Data Products Program

advantages to such an approach would be to make the validation software more data-driven, and the process of providing content for the tables subject to configuration and quality control. *We recommend the construction and maintenance of one or more configured reference tables to support the process of metadata validation.*

## 8 Summary of System Change Requests

The change requests that need to be filed to implement the recommendations in this STR include the following:

1. If the keywords described in the Tables are not present in datasets for all instruments with imaging modes, they must be added.
2. The values of keywords described in this document must be validated according to the criticality criteria described in Section 1.
3. Keyword validation must be tracked either by individual keywords, or by category as defined by the titles of sections 2 through 6 of this document. Validation values are one of:
  - *pass*, for values that have been examined and are believed to be correct,
  - *fail*, for values that have been examined and are believed *not* to be correct, or
  - *unknown*, if validation has not been performed.
4. Keyword values must be remediated according to the criticality criteria described in Sect. 2.
5. The Customer Team will create tables of instrument configuration information that will aid in the keyword validation and remediation process. These tables will be delivered to the Operations staff, along with documentation for how they will be used. The tables will constitute part of the configured (re: controlled) E2E system. The tables to be delivered include the following:
  - WCS keywords that describe the nominal plate scale, tangent point, and related information
  - Filter names, bandpasses, and other identifying information
  - Look-up tables or functions for computing Local Siderial Time, given exposure date and time

## 9 References

- Greisen, E. W. & Calabretta, M. R. 2002, *Representation of World Coordinates in FITS*, AA, 395, 1061
- Calabretta, M. R., & Greisen, E. W. 2002, *Representation of Celestial Coordinates in FITS*, AA, 395, 1077
- IAUFWG 2005, Definition of the Flexible Image Transport System (FITS), Version 2.1b, available at <http://fits.gsfc.nasa.gov/standard21b.html>.
- Shaw, R. A., & Miller, C. 2007, *Dataset Identifiers*, DPP System Technical Report 2007-02 (Tucson: National Optical Astronomy Observatory)
- Shaw, R. A., ed. 2007, *NOAO Data Handbook*, (Version 1: Tucson: National Optical Astronomy Observatory)
- Valdes, F. 2000, NOAO Image Data Structure Definitions, available at <http://iraf.noao.edu/projects/ccdmosaic/imagedef/imagedef.html>.