

Forensic Data Quality Assurance

Richard A. Shaw, Robert L. Seaman, Sonya Lowry, Howard Lanning

National Optical Astronomy Observatory, Tucson, AZ, USA

Abstract. Accurate metadata are essential for the scientifically meaningful interpretation of data products. Yet metadata from legacy NOAO data-taking systems are often incomplete and sometimes inaccurate. A new Data Quality Assurance (DQA) system for NOAO is proposed to ensure the best possible data and metadata accuracy, integrity, and integrability with the Virtual Observatory.

1. Introduction

The veracity of the metadata that describe scientific data, such as world coordinate systems, time-stamps, photometric calibrations, instrument configurations, etc., is often taken for granted by astronomers. Yet accurate metadata are critical for the scientifically meaningful representation of data products and for their proper interpretation by researchers. These metadata are for the most part generated contemporaneously with the data by various subsystems in the observing environment, and are persisted in the form of keyword/value pairs in the science FITS files. We describe a new Data Quality Assurance (DQA) System for the NOAO Science Archive (NSA) that will assure compliance with the science and operational requirements for data and metadata accuracy, integrity, and integrability with the Virtual Observatory (VO).

The need for a DQA system at NOAO has two main drivers that distinguish the NSA from many well established archives. First, the NSA is obligated to ingest data products from nearly three dozen legacy instruments and data taking systems (see the data transport diagram in Fig. 1 of Barg et al. 2007) that were not designed with archives and VO end users in mind. Core metadata are represented in a variety of ways in these data headers, with varying levels of integrity, accuracy, and completeness; key metadata are sometimes missing entirely. Second, even if upgrades to the data-taking systems could be made, we would still be faced with the challenge of improving the quality of metadata for millions of datasets that have already accumulated over the past few years.

2. DQA Objectives

The goals of data quality assurance are multifaceted (Radziwill 2007), but those of greatest concern here are to bring all critical metadata to a common and well defined standard of quality and completeness, and to establish a robust base upon which other value-adding processes (such as pipelines) can build. It is also important to enabling a uniform and consistent view for all data resources, including that for the public interface to the NSA, and for private interfaces

to sub-systems within the NOAO end-to-end data management system. These goals will be achieved by assessing and recording the validity of the essential metadata, repairing faulty metadata where possible and feasible, adding additional metadata that were not available in the observing environment, and regularizing the form and format of the metadata.

The scope of DQA will initially be limited to core metadata that are most relevant to the VO user of the archive who may have little or no familiarity with NOAO instruments, or ground-based Optical/IR data. These metadata consist of those that describe provenance, plus *coverages* in the VO sense (Rots 2007):

- **Provenance** identifies the originator of the data, the configuration of the system(s) used to obtain the data, the nature and pedigree of any subsequent processing of the data, and unique identifiers for the dataset within the data publisher’s domain (such as those for IVOA or ADS).
- **Spatial coverage** describes the footprint of the instrument aperture(s) on the sky, the mapping of detector coordinates to world coordinates (FITS 2008), and spatial resolution information.
- **Spectral, or bandpass coverage** describes the attributes of the effective bandpass of the data, including bandwidth and spectral resolution.
- **Temporal coverage** describes the interval(s) of time during which photons were collected, including the time system, time stamp(s) denoting the start time(s), and duration(s).
- **Brightness coverage** describes the dynamic range of the observation, and ordinarily includes the conversion of instrumental brightness units to physical units, the background level, and saturation and linearity limits.

Over time it will be possible to expand the scope of DQA to additional metadata, and to measure defect rates for metadata that originate in the observing environments, which will provide a basis for diagnosing problems with instruments and their data taking systems.

3. DQA System Objectives

It is perhaps important here to define certain key terms with more precision: to *validate* is to determine whether required metadata are present and accurate. If a metadatum fails validation, it is sometimes possible to *remediate*, which is to assure compliance with NOAO standards by correcting, regularizing, and adding metadata through standard and controlled processes. By *correct* we mean replace flawed or incomplete metadata accurate values; to *regularize* is to assure that the content conforms to NOAO and FITS standards for type, units, and syntax.

The system that implements DQA functionality must also meet certain objectives. In particular, the quality assurance activities must be:

Systematic DQA processing is applied consistently to all datasets.

Controlled Use controlled/configured software and rules to perform the DQA processing, rather than ad-hoc fixes. Minimizing human intervention promotes homogeneity in metadata quality—this includes capturing only well defined and structured data from the observing environment, and resisting

temptations such as harvesting free-form observer logs which are inhomogeneous in content and scope, and are too often wrong.

Traceable Updates to metadata must be auditable to assure compliance with data management policies, to assure that the system is robust against unplanned down-time, and to provide a means for faulty updates to be identified and repaired. Traceability also enables constructing views of the system with thresholds on quality metrics.

Automated The system must be automated to the extent feasible, to minimize both operations costs and inconsistencies introduced by human interaction. Automation also means the processes are repeatable, and changes to the DQA system can be measured directly and unambiguously.

Extensible Not all DQA needs or strategies can be anticipated in advance, and the implementation of the DQA system itself may have flaws that must be addressed. NOAO data headers for newer instruments are well defined and relatively complete, but others require extensive remediation. Therefore DQA activities will start with the newest instruments with high user demand, and may later be extended to more problematic cases.

Adaptable Enhancements and corrections will rely far more on changes to rules and configuration files than on changes to control logic. Thus many important changes can be implemented and tested without a costly cycle of source code updates and releases.

At the conclusion of DQA processing each metadatum will have an associated status, shown in Table 1, which tags the end result. The initial copy of the data and metadata, as captured in the observing environment, are never deleted, so it will always be possible to repeat any aspect of DQA processing.

Table 1. Metadatum Status Following DQA Processing

Status	Meaning
ADD	Both a missing keyword and a correct value must be added.
DELETE	Keyword should be redacted, such as the deprecated EPOCH, or WCS keywords from images of internal calibration sources.
FAIL	Metadatum value failed validation check; no remediation possible.
PASS	Metadatum value passed validation; no remediation necessary.
UPDATE	Metadatum value remediated; replacement value was derived.
UNKNOWN	[Default] Metadatum has not been validated.

4. DQA System Context

The evaluation of science data quality, and the use of processes for metadata remediation, are not limited to any particular sub-system in the NOAO Data Management System. Indeed, even pipeline processing can be thought of as a process that creates or improves core metadata. Nor are these evaluations isolated in time, since the scope of remediation will likely grow. But it is important

that the responsibility for validating and remediating any given metadatum, and the circumstances under which it occurs, be unambiguous. DQA processing may be performed at any time after the raw data product is ingested, but preferably prior to pipeline calibration; it may also be initiated any time the validation or remediation rules are introduced or modified. All critical metadata will be stored in the NSA database, including the initial value from the raw data ingestion, the current value (as updated by DQA or pipeline processing), and the status information described in §3. Actual updates to data headers will be deferred until a request is received for download. This “lazy” instantiation of updated headers ensures that DQA processing can be executed at any time without conflicting with other data management activities, that the operational cost of header updates is not incurred unless and until a dataset is requested, and that the best, most correct metadata are always presented to users.

5. Remediation Strategies

For many NOAO instruments, critical metadata are recorded in the FITS header, but with non-standard keyword names. It is easy in such cases to construct a rule to remap keyword aliases. We have also developed heuristics, for example to identify discrepancies between exposure start times and the file write timestamps from the data taking system (modulo the exposure time). Many of the 19 imagers in the NOAO suite of instruments do not record full WCS information in the data headers, and even basic telescope pointing information can be surprisingly inaccurate. In these cases the automated astrometry solver described by Hogg, et al. (2008) shows high promise for generating accurate WCS metadata for all our instruments, particularly since we know the plate scales to high accuracy. Finally, we are developing an heuristic to use image histograms to distinguish between science targets and internal calibration exposures. This is important for recognizing proprietary (science) data, access to which is governed by the observatory data-rights policy.

Acknowledgments. We were saddened by the untimely death of our colleague, Howard Lanning, in late 2007. He was a constant source of calm, persistence, and devotion to scientific quality.

References

- Barg, I., Seaman, R., Lanning, H., Smith, R. C., & Saavedra, N. 2007, in ASP Conf. Ser. 376, ADASS XVI, ed. R. A. Shaw, F. Hill, & D. J. Bell (San Francisco: ASP), 515
- FITS 2008, Definition of the Flexible Image Transport System, IAU FITS Working Group (Version 3.0; Greenbelt, MD: FITS Support Office, NASA/GSFC); available on-line at http://fits.gsfc.nasa.gov/fits_standard.html
- Hogg, D. W., et al. 2008, in ASP Conf. Ser. 394, ADASS XVII, ed. R. W. Argyle, P. S. Bunclark, & J. R. Lewis (San Francisco: ASP), 27
- Radziwill, N. M. 2007, in ASP Conf. Ser. 376, ADASS XVI, ed. R. A. Shaw, F. Hill, & D. J. Bell (San Francisco: ASP), 363
- Rots, A. 2007, Space-Time Coordinate (STC) Metadata Model, IVOA Note 5, available on-line at <http://www.ivoa.net/Documents/latest/STC-Model.html>