

Dataset Identifiers

Richard A. Shaw (DPP, NOAO/Tucson) and Christopher Miller (DPP, NOAO/La Serena)

9 May 2007

Dataset identifiers must be established for each data product that is archived by DPP. In general there may be more than one way to refer to a dataset within the DPP E2E system, but from the point of view of an external user the identifiers must satisfy a number of criteria, among them permanence, uniqueness, and the ability to refer to specific products from external publications. This document reviews community standards and conventions, and defines a simple nomenclature from which identifiers will be created for those products that are visible to the external user. This scheme accommodates calibration reference files, datasets that are contributed from the community for publication in the archive, and anticipates extensions to the nomenclature for new types of data products.

1 Introduction

In astronomy, as with most fields of science, there is a strong tradition of citing the particular observations that support the conclusions drawn in a publication. Such references give both credibility to the author, and support the essential tenet in science that results must be verifiable by independent investigators. With the advent of permanent, digital archives of astronomical data it has become possible not only to refer with unprecedented precision to data that were used, but for other researchers to use the very same data for a variety of purposes, including validating prior research results. New standards and technologies that are being developed in the context of the Virtual Observatory (VO), coupled with conventions being developed by the major publishers of the refereed astronomical literature (AAS 2006), are making it possible for readers to retrieve supporting datasets directly from links embedded within papers. Not surprisingly, there are additional, emergent benefits associated with these new technologies and services. One is the new service provided by the Astrophysics Data System (ADS) that allows publishers of astronomical data (usually the major archive centers) to harvest information to determine which datasets within their holdings have been cited (Accomazzi 2004); another is to provide links from those datasets back to the literature (Accomazzi, et al. 2007). Such a capability is extremely useful to researchers, and the number of references to an archived dataset is a useful measure of its value.

Although no detailed requirements on dataset identifiers have yet been established for the next generation of the NOAO Science Archive (NSA, v3), it is critical that identifiers satisfy the expectations of the external community, and are compatible with the requirements of both the VO and the publishers of the refereed literature. It should be noted that the identifiers described in this document refer only to identifiers that are used for accessing NSA data products from external web services and users. It is possible that a separate identifier will be used by systems internal to DPP, but in this case the relevant DPP systems must assume responsibility for the mapping

Data Products Program

between internal and external identifiers. Satisfying these external requirements provides an opportunity to organize the data holdings of the NSA in new ways that aid the user community.

2 Dataset Identifiers in the Community

There are a variety of attributes that are necessary for viable dataset identifiers, including those mentioned above. The focus in this section is on the community expectations, and specifically the IVOA standards for referencing datasets. Such compliance will assure that all NSA datasets will be accessible with VO-enabled tools that invoke IVOA protocols to search for data of interest to users. Similarly, it is important that datasets be accessible from professional journals that reference them. These two needs can be met with the protocols described below, which are similar but not identical; it is possible to define a nomenclature that bridges the differences between the two, as explained in §3.

2.1 IVOA Identifiers

Plante et al. (2006) describe the syntax for *resource identifiers* that will be required for the International Virtual Observatory Alliance (IVOA), which is based upon the internet standard for *Uniform Resource Identifiers* (URIs: see Berners-Lee, Fielding, & Masinter 2005). Note that a *resource* in this context is quite general, and can refer to documents, data files, web services, etc. For the present purpose we consider resources in a more restricted sense, i.e., either to data files or collections of data objects offered through the NSA, or to an interface or service that provides access to these products. The definition for IVOA resource identifiers is, in essence, a mechanism to provide a persistent, globally unique reference to any resource on the web that can be described by generic resource metadata defined by the IVOA, using the specified syntax. The URI form of a resource identifier is:

```
ivo://AuthorityId/ResourceKey
```

The *authority identifier* is a compact string that identifies the *naming authority* that creates IVOA-compliant identifiers for the resources it registers. The authority identifier defines a globally unique name-space that is controlled by that naming authority. The second part of the resource identifier is the *resource key*, which is the name for a localized resource that is unique within the name-space of the `AuthorityID`. The `ResourceKey` is composed of one or more *segments*, delimited by slashes (“/”), and that may contain a very limited set of non-alpha-numeric characters, including the period, dash, and underscore. The careful use of segments and the non-alpha-numeric characters provides an opportunity to create meaningful (i.e., human-interpretable) identifiers within the NOAO name-space (see §3.1).

It is important to bear in mind the distinction between an IVOA identifier and the object to which it refers. Given an IVOA identifier, all that is required is that a user *be able* to retrieve the named dataset—i.e., the URI *identifies* the dataset uniquely, but it does not necessarily provide a location nor specify a protocol by which it can be accessed. Users will commonly encounter IVOA identifiers in an electronic journal, or as the result of a search for data matching a set of criteria. Upon de-referencing the identifier (e.g., by following a link in a web browser), we expect the user to be presented with a web service or other mechanism that will provide, e.g., a URL to retrieve the named dataset.

The IVOA standards do not define a *dataset*; rather, the definition is left to the data providers. Judging by data products that are offered from major, existing archive centers, it is common for

the datasets themselves to have structure (e.g., to be composed of multiple atomic pieces). Indeed, the name *dataset* has been widely used to refer both to a single disk file containing data, and to a collection of intimately connected data; often the two are really one in the same since, e.g., it is common to package multiple science data frames within a single multi-extension FITS (MEF) file. It is the responsibility of the DPP Customer Team to define and document the organization of the data products that are published in the NSA, and how to present this organization to the user.

2.2 ADS Identifiers

The ADS, in collaboration with the University of Chicago Press (the current publisher of the refereed journals sponsored by the American Astronomical Society) and the IVOA, has developed a method for authors to reference datasets in manuscripts. Authors using the L^AT_EX mark-up language can use a macro like the following to tag the datasets:

```
\dataset{ADS/FacilityID#PrivateID}
```

The items in bold within curly braces are the `AuthorityID`, followed by a *facility identifier*, then a *private identifier* that is a surrogate for the resource specification. The `AuthorityID` in this case is the ADS, which is assuming the responsibility for mapping between the dataset citation in the journal to the actual resource on the web. In particular, ADS has taken on a number of responsibilities for linking between journals and datasets (Accomazzi 2004), including:

- hosting a central dataset identifier service to validate identifiers and associated URLs (i.e., verifying that the identifier can be de-referenced, and actually provides access to the named resource);
- providing the URIs back to the publisher for constructing live links in electronic journals;
- providing a service to data centers to harvest literature references to datasets they host.

The `FacilityID` needs to be established in collaboration with the ADS, and the `PrivateID` is established within DPP. The `PrivateID` will have a fairly direct mapping to the actual names of datasets (e.g., filenames or collections thereof) that are published by the NSA, and will be identical to the IVOA `ResourceKey`.

3 A Nomenclature for the NOAO Science Archive

Building on the community expectations for dataset identifiers, it is clear that developing a viable nomenclature for data that are published by the NSA amounts largely to specifying the `AuthorityID` and the `ResourceKey` components of an IVOA identifier; this should be done in a way that is sufficiently general and extensible to accommodate all data resources that are offered now or might be offered in the future. The following subsections describe the new nomenclature in detail, including a specific example for a typical dataset.

Data Products Program

3.1 Dataset Identifier Anatomy

3.1.1 AuthorityID

The `AuthorityId` should reflect the role that NOAO/DPP plays in publishing datasets from the facilities it operates, as well as from partner institutions. For data products that originate with one unique observatory, the corresponding URI would be of the form:

```
ivo://NOAO.<ObservatoryId>/ResourceKey
```

where *ObservatoryId* is a string of up to four alphabetic characters. For example, the `AuthorityId` for datasets provided by the WIYN Observatory would be `NOAO.WIYN`, which reflects both the originating observatory and the publisher of the data. Similar namespaces will be reserved for KPNO, CTIO, SOAR, SMRT (for the SMARTS consortium) and so on for other partners as the need arises. The string used for `AuthorityId` will be identical to the `FacilityId` used for the ADS identifier. For data products that are assembled from multiple observatories (e.g., for catalogs), the `AuthorityId` string will be simply `NOAO`.

3.1.2 ResourceKey

The `ResourceKey` is the primary, deep reference to a dataset within the context of the archive's holdings. It creates an opportunity to tag the holdings in a way that facilitates referencing datasets concisely, whether in the literature or on the web. The `ResourceKey` will be composed of the following elements:

```
ResourceKey = ProjectId/ArchiveKey
```

where both of the components are compact alpha-numeric strings, separated by a slash, that together provide a unique reference to a dataset. *ProjectId* identifies project that is intellectually responsible for producing the data product; it is ordinarily derived from the observing proposal ID. Note that data not associated with a proposal (for historical or other reasons) will be treated as a special case: a `ProjectId` will have to be created, along with *ArchiveKeys* for all of the data products that are published for the project. The *ArchiveKey* is an alpha-numeric string that uniquely identifies the dataset within the scope of the project, and is a simple serial number, or (for compactness) a serial number that is mapped to base-36. When the `ArchiveKey` is absent, the `ResourceKey` implicitly refers to the entire project collection. The complete `ResourceKey` string will be stored in FITS file primary headers, as the value of the `VO_IDENT` keyword. To conform to the FITS standard, the length of this string must not exceed 64 characters.

3.1.3 ADS Identifiers

Identifiers that are compatible with ADS requirements will be fashioned directly from the IVOA dataset identifier:

```
ADS/NOAO.ObservatoryId#ProjectId[/ArchiveKey]
```

The entire ADS identifier will be stored in FITS file primary headers, as the value of the `DS_IDENT` keyword. Again, it is permissible to reference an entire data collection in a journal paper by omitting the `ArchiveKey` portion of the identifier. This requires that a DPP/NSA service be implemented that provides access to all members of the data collection.

4 Application to Data Products

The nomenclature scheme for *raw* datasets that are generated by NOAO and partner observing facilities follows directly from the preceding discussion. The nomenclature for other data products is discussed in the following subsections.

4.1 *Processed Data Products*

Higher-level data products that are produced by data reduction pipelines will include reduced and calibrated science images, data quality masks, metadata, thumbnail images, as well as ancillary data such as source lists. Each data file that is ingested into NSA will be assigned an identifier, even if the products are not initially offered to VO users. These products will share the same `ProjectID` as their parent datasets; the `ArchiveKey` will be a serial number that is unique within this project namespace.

4.2 *Calibration Reference Files*

Processed calibration data products are generated by instrument-specific pipelines (as well as through other means), usually by combining calibration images or by deriving characterizations of data. These data are non-proprietary, and may be used for the purpose of calibrating science data that were obtained for other programs. These data products will be assigned a dataset identifier upon ingest into the NSA: a special project identifier “**Ca1DB**” will be used, and the `ArchiveKey` will be a serial number that is unique within this project namespace.

4.3 *Contributed Datasets*

Datasets that are provided by external parties, such as survey teams (and currently published via v2 of the NOAO Science Archive) will at some point be offered through versions 3 and higher of the NSA. The scheme outlined above extends naturally to contributed datasets, except that a unique `ProjectID` will be generated by the Customer Team. `ProjectIDs` for the datasets that have been contributed through the NOAO Surveys Program prior to this writing are given in the Appendix; additional `ProjectIDs` will be provided by the DPP Customer Team as the need arises.

4.4 *Extensions to the Nomenclature*

There will undoubtedly be new types of data products offered through the NSA. New types of science products may include time-series (e.g., light curves) and catalogs. Such products could be derived from several or many datasets (as defined here), so that it may be more sensible to invent new `ProjectIDs` than to require that they map to existing projects.

Other kinds of products, such as contemporaneous images from all-sky cameras or weather satellites, pertain to conditions for a particular night of observing rather than specific datasets. For these ancillary data products the identifiers will be grouped with special `ProjectIDs`, indicating their association to the calibration and data quality activities for the observatory and instrument, and to the non-proprietary nature of these data products. If/when observing logs become available as archived data products, they may be grouped with the observing program that generated them.

Data Products Program

4.5 Versioning

This nomenclature scheme supports versions of data products indirectly, in that a new `ArchiveKey` is generated for every new product ingested by the NSA. However, it is always required that version information be present in the metadata (including the file header keywords, if relevant) for a dataset, even if it is the only version available. The detailed syntax for how the version information is to be represented in metadata is deferred until the need arises, and until a programmatic/scientific decision has been reached on whether multiple versions of data products will be offered.

4.6 Mapping to Files

Dataset identifiers must be mapped to physical files once they are provided to the user for some purpose, such as manipulating content on the VO portal or in response to a retrieval request. It is a straightforward exercise for (the relevant) DPP systems to parse a `ResourceKey`, derive the internal reference to one or more disk files, and take the appropriate action. The focus here is on the transformation to the names of physical files to which a user has access, which can be accomplished in a few ways. By default, the recipe consists of mapping the `ResourceKey`¹ to the file name and appending the appropriate file extension (usually `.fits`) to denote the data format. Note that the forward slash must be mapped because it is a reserved character on most file systems; for this purpose the underscore (`'_'`) will be used instead.

As an example, consider one file that was generated by NOAO observing program 2005B-0045 on the night of 28-July-2006 at the CTIO 4-m telescope. The DTS assigned the file name: `ct654996.fits`. The `ProjectID` is taken directly from the observing program ID. The `ArchiveKey` is simply the value of the `DTNSANAM` keyword, with the numeric portion converted to base-36, and without the file extension: `ctE1EC`. The fully qualified identifiers of interest for this dataset are:

```
NOAO Observing Program ID: 2005B-0045
DTS file name: ct654996.fits
IVOA URI: ivo://NOAO.CTIO/2005B-0045/ctE1EC
ADS Identifier: ADS/NOAO.CTIO#2005B-0045/ctE1EC
Default File Name: 2005B-0045_ctE1EC.fits
```

It is also possible to generate alternative file names, according to a user preference. For instance, the filename that was supplied at the telescope when the data were obtained may be found in the image header as the value of the `DTACQNAM` keyword. Other schemes, involving the date, telescope, instrument, and other metadata that are found in the header could also be constructed. A tool should be created to apply one (of a few) filename mapping schemes, such that users could

¹ Note that the `ArchiveKey` is required when mapping to file names.

invoke it as the data are staged at the NSA, or alternatively so that users can re-name their data files after retrieval.

5 Implementation

5.1 *Activities*

The following activities need to be carried out to implement dataset identifiers; most will involve E2E system change requests. The order of the activities listed below is not significant, but see §5.2 on priorities. Details for implementing the ADS-related activities may be found in Accomazzi (2004).

- [ADS] Register the NSA with the ADS. Re-register the NSA with the IVOA identifier(s) specified in this document.
- [ADS] Create a site profile for NOAO data products that can be accessed by the ADS.
- [ADS] Build and deploy a dataset verification service.
- [ADS] Implement a service for harvesting data references from the ADS.
- Construct a (configured) document that describes the data products that are published in NSA. The descriptions will be specific to each telescope/instrument and operating mode where the data products differ, and must include a physical description of what is represented and the mapping onto the file structure.
- Implement software that creates dataset identifiers, following the scheme described in this document.
- Create or modify archive ingest software to: insert the dataset identifier keywords into the headers of FITS files that will be ingested, and include the IVOA and ADS identifiers in the archive database.
- Define a file that would be constructed for each delivered dataset that describes for a user the content of each delivered file.
- Create a tool or service to map from NSA file names to names that are more meaningful to an archive user. This should include the possibility of using these alternative names when users download data from the NSA, and for re-naming the files in the user's storage area once the data have been down-loaded.

5.2 *Priorities*

The essential requirements that dataset identifiers be unique and permanent, along with the programmatic desires that researchers make effective and visible scientific use of the NSA, make implementation of substantial parts of the nomenclature scheme the highest priority. At a minimum, a file nomenclature scheme that meets the requirements specified in this document must be in place at the time when calibrated data from any instrument (and as produced by DPP pipelines) are first made available through the NSA. It is highly desirable for the nomenclature scheme to be in place by the time any non-proprietary raw data are available through the NSA.

6 References

AAS 2006, *AAS Data Set Linking* (Chicago: University of Chicago Press), available at <http://www.journals.uchicago.edu/AAS/datasets/>

Accomazzi, A. 2004, *ADS Dataset Verification and Resolution Services*, available at <http://vo.ads.harvard.edu/dv/>

Accomazzi, A., Eichhorn, G., & Rots, A. 2007, in ASP Conf. Ser., *Astronomical Data Analysis Software and Systems XVI*, ed. R. A. Shaw, F. Hill, & D. Bell (San Francisco: ASP), in press

Berners-Lee, T., Fielding, R., Masinter, L. 2005, *Uniform Resource Identifier (URI): Generic Syntax*, RFC3986, available at: <http://gbiv.com/protocols/uri/rfc/rfc3986.html>

Plante, R., Linde, T., Williams, R., & Noddle, K. 2006, *IVOA Identifiers*, IVOA Proposed Recommendation 2006-08-22, Version 1.11, available at <http://www.ivoa.net/Documents/PR/Identifiers/Identifiers-20060822.html>

7 Appendix: ProjectIDs for Contributed Datasets

The following table defines the `ProjectID` for datasets that have been contributed by the NOAO Survey programs:

| ProjectID | Survey Program |
|------------------|---|
| CHMP | ChaMPlane: Measuring the Faint X-ray Binary and Stellar X-ray Content of the Galaxy |
| DPES | Deep Ecliptic Survey |
| DPLS | Deep Lens Survey |
| DPRS | Deep-Range survey |
| NDWFS | NOAO Deep Wide-Field Survey |
| FLMX | FLAMEX: FLAMINGOS Extragalactic Survey |
| FLS | First Look Survey |
| FPPVS | Fundamental Plane Peculiar Velocity Survey of Rich Clusters |
| FSVS | Faint Sky Variability Survey |
| LGSC | Resolved Stellar Content of Local Group Galaxies Currently Forming Stars |
| SINGG | Survey for Ionization in Neutral Gas Galaxies |
| SMCHO | Next-Generation Microlensing Survey of the LMC (a.k.a. SuperMACHO) |
| WPRJ | The w-Project: Measuring the Equation of State of the Universe |
| ZBOOT | z-Band Observations of the NOAO Deep Wide-Field Survey Bootes Field |

Although there is no formal restriction on the length of the `ProjectID` names, they should be limited to about 6 characters in the interest of compactness.