Astronomical Data Analysis Software and Systems XIII
ASP Conference Series, Vol. 314, 2004
F. Ochsenbein, M. Allen, and D. Egret, eds.

Publishing Links to Astronomical Data On-line

Alberto Accomazzi, Günther Eichhorn

Harvard-Smithsonian Center for Astrophysics

Abstract. We discuss the design and implementation of a scheme enabling authors to refer and link to on-line datasets available from astronomical archives. This will provide the readers of electronic papers with direct access to the data discussed therein. The software tools used to create and maintain links from published papers to the datasets make use of Web-Services-based technology. The system has been designed in collaboration with the NASA Astrophysics Data Centers, the American Astronomical Society, and the University of Chicago Press, and will be maintained by the NASA Astrophysics Data System. More information about this project is available at: *http://vo.ads.harvard.edu/dv*.

1. Introduction

This paper describes the Dataset Verification and Linking efforts underway among the NASA Archives and Data Centers, the American Astronomical Society (AAS), and the University of Chicago Press (UCP, publisher of ApJ, AJ and PASP). This activity has taken place under the auspices and guidance of the NASA Astrophysics Data Centers Executive Council (ADEC), and aims at fulfilling the promise of further integrating the astronomical literature and the on-line data it is based upon.

The NASA Astrophysics Data System (ADS) is developing the tools needed by publishers and users at large for both dataset verification and linking through stable, top-level services that can be maintained for the foreseeable future. Links created to datasets from on-line manuscripts will always refer to a dataset via a URI created using a well-defined identifier, and the URI will be turned into one or more URLs in real-time by a central resolver provided by the ADS. This will provide a high level of reliability and persistence to the links, as well as providing an upgrade path into any future Virtual Observatory (VO) efforts in this direction. Dataset citation, verification and linking will work as follows:

- Astronomy data centers and archives will start attaching permanent dataset identifiers to the data they distribute.
- Astronomers will write papers referencing the dataset they have used in their research. As per the instructions given to them by the AAS, they will start using the appropriate markup to identify datasets in the papers they publish.
- During the publishing pipeline, UCP will extract the identifiers and send a query to a central dataset identifier service (hosted by the ADS) to find out if (a) the dataset is valid and (b) a URL can be associated to it.

- The central dataset identifier verification service will query a number of (relevant) datacenters using its own protocol, will cache the results, and will return a status flag indicating if a dataset is known or not.
- For the dataset identifiers that are known, URLs can be built by using the base URL of a dataset identifier resolver and the dataset identifier itself, e.g. *http://vo.ads.harvard.edu/dv/DataResolver.cgi?ADS/Sa.CXO#15.* If the verification is successful, UCP will include such a URL in its on-line article.
- When the article goes on-line, a user clicking on the link associated with the dataset will be taken initially to the URL above. What happens next depends on whether the ADS has one or more datacenters claiming to have data relative to this dataset (there could even be different mirror sites for a given data center). If only one final URL is available for the dataset in question, the resolver will simply forward the user to it. If more than a single URL is available, a simple menu listing all the information we have about the available links will be displayed.

ADS will take the responsibility of maintaining services that are aware of all relevant datacenters that may have datasets available on-line, and datacenters profiles indicating which datasets are available from each of them.

2. Dataset Identifiers

In order to allow easy integration of this effort in the emerging VO framework, the ADEC has decided to adopt a syntax for the dataset identifiers which is consistent with the current International Virtual Observatory Alliance (IVOA) Dataset Identifier draft (Plante et al 2003). This adoption will facilitate integration of these identifiers and the tools that manipulate them in the VO.

2.1. IVOA Identifiers

According to the IVOA Identifiers Draft, the general URI format for an individual identifier is a string of the kind: *ivo://AuthorityId/ResourceKey#PrivateId*. While we refer the reader to the draft for a full explanation of the syntax, a few things are worth pointing out:

- Use of the *ivo://* scheme denotes the fact that the rest of the identifier should be interpreted as a string abiding by the IVOA Identifiers specification, and that the identifier and the resource it refers to have been registered with an IVOA-compliant registry.
- *AuthorityId* is a naming authority registered within the IVOA community; the use of this string within the identifier establishes a namespace within which the rest of the identifier can be considered unique. In general, the *AuthorityId* does not need to correspond to a specific institution but rather to an entity that has been granted use of the namespace.
- *ResourceKey* is a name for a resource that is unique within the namespace estabilished by the *AuthorityId*. In general it will correspond to a unique resource made available to the VO by or on behalf of the *AuthorityId*. A typical example of a *ResourceKey* in this context is a data collection generated by a particular project or mission.

- *PrivateId* represents a unique string within the *ResourceKey* and it denotes a particular dataset belonging to the collection.

2.2. Using Dataset Identifiers in the Literature

Given the fact that much of the VO infrastructure is still under design and development, the ADEC has decided on a specific recommendation for referring to dataset identifiers in the astronomical literature. The general form of these identifiers is: *ADS/FacilityId#PrivateId*. Comparing these identifiers with the general IVOA syntax we can make the following observations:

- No protocol scheme has been specified. This is due to the fact that until IVOA-compliant registries are available, and *AuthorityId*s can be established by them, it would be incorrect to claim that these identifiers are in fact IVOA compliant. However, it is to be expected that these identifiers can be resolved as IVOA identifiers in the not too distant future by a simple syntactic operation.
- The *AuthorityId* string "ADS" has been specified. This simply recognizes the current role of the ADS in managing the namespace used for these identifiers, in the absence of a community-wide namespace granting authority. It does not suggest nor imply that the ADS controls or manages the dataset itself.
- The *ResourceKey* token will be interpreted as a Facility. An ever-growing list of facilities is maintained by the ADS. Data centers should contact the ADS should they need to register new entries.
- The *PrivateId* string can be anything that the data center desires, with the provision that the identifier string as a whole should abide by the general syntax of a URI, as required by the IVOA identifiers specification.

3. Generating Dataset Identifiers

All Data Centers and Archives which provide public access to their data should structure their databases and interfaces so that when a particular dataset is released to the public, it is uniquely tagged by an identifier ID created as discussed above. Users who download such a dataset should be made aware of the identifier associated with it and how it should be referenced in the published literature. In order for a datacenter to ensure that the identifiers it is generating comply with the syntax endorsed by the ADEC, the following must occur:

1. The identifier is in the form *ADS/FacilityId#PrivateID*
2. The *FacilityId* has been registered with the ADS and is listed in the table of known facilities
3. The *PrivateId* is a unique identifier within the *FacilityId*, and its association with the dataset will not change.
4. A profile for the datacenter has been registered with the ADS, and *FacilityId* has been listed as one of the resources that the center has data for.
5. The datacenter provides a dataset verification service which will be used to verify the validity and location of identifiers published in the literature.

Once a datacenter has published a dataset ID, it should provide access to it. This should be a human-readable page on its web server displaying the dataset's relevant metadata and offers the user the option to download the dataset itself in some form or fashion. It is left up to the datacenter to decide what to do if and when a revised version of a particular dataset is published. In general, however, it is understood that access to the latest revision of a dataset should be an option if not the default.

4. Providing Data Verification Capabilities

In order to promote an open framework that can be used for the distributed verification of dataset identifiers across data centers, the ADEC ITWG (Interoperability Technical Working Group) has created the specification for a SOAP-based web service. The corresponding WSDL file can be used to generate client and server interfaces to the service. Each datacenter providing data verification services should provide and maintain a service that abides by this specification.

In order for the ADS to coordinate the verification and linking of dataset identifiers to the appropriate datacenters, it is necessary for the datacenters to provide some basic metadata about its data holdings and services. While it is expected that the appropriate metadata will one day be made available by a public VO registry, its format and access methods are at this time not available. As an intermediate solution to the problem, we require that the data centers maintain a simple profile which will provide the ADS with the necessary metadata to maintain a central verification service that fans out queries to the appropriate datacenters (during the verification phase) and links to the individual datasets (during the link resolution phase).

The data center profile is a simple XML document that lists the data center name and description, the name and email address of the person responsible for the maintenance of the profile, the URL of the web service to be used for dataset verification, and the list of facilities that the datacenter has data for. The central verifier service will only attempt to verify and link a dataset identifier with a datacenter if its profile indicates that the datacenter archives the appropriate data collection.

To facilitate the deployment of verification services, the ADS also developed a PERL toolkit that greatly simplifies the creation of a compliant web service. Among other things, by defining a few variables and installing a simple CGI script based on this toolkit a system manager will be able to automatically define his/her site's profile described above. For more information, please see the project's description available at *http://vo.ads.harvard.edu/dv.*

Acknowledgments. The NASA Astrophysics Data System is funded by NASA Grant NCC5-189.

References

Plante R. et al. 2003, IVOA Identifiers Working Draft v.0.2 (30 September 2003), *http://www.ivoa.net/Documents/WD/Identifiers/WD-IDs.html*