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

Resource Metadata for the Virtual Observatory

R. J. Hanisch, G. Greene

Space Telescope Science Institute

A. E. Linde, K. T. Noddle

University of Leicester

R. L. Plante

National Center for Supercomputing Applications, University of Illinois Urbana-Champaign

A. M. S. Richards

Jodrell Bank Observatory

E. C. Auden

Mullard Space Science Laboratory

W. O'Mullane

The Johns Hopkins University

Abstract. The location and access methods of astronomical resources (catalogs, observation logs, and data archives) and associated computational services (e.g., data processing pipelines, source extraction services, theoretical simulations) in the Virtual Observatory will be determined by querying dynamic resource registries. These registries function as a sort of yellow-pages, providing descriptive information (metadata) about the resources in order to locate information and services in response to user queries. The metadata also needs to describe the provenance of the information, provide some indication of the data quality, quantity, and type, and guide users to information appropriate to their needs (i.e., research-oriented data archives vs. educational resources).

1. The Role of Metadata in the Virtual Observatory

In order to make it easy for astronomy information services to participate in the VO, we propose a system for metadata management based on a hierarchy of descriptive schemas. At the top level we require a minimum amount of information, sufficient primarily to note the existence of a resource and to describe who is responsible for it. At lower levels, the metadata are more extensive and complex, allowing for the description of query syntax, access protocols, and usage policies.

A *resource* is a general term referring to any VO entity that can be described and which can be given a name and unique identifier. Just about anything can be a resource: it can be an abstract idea, such as sky coverage or an instrumental setup, or it can be fairly concrete, like an organization or a data collection. This definition is consistent with its use in the general Web community as "anything that has an identity" (Berners-Lee et al. 1998). We expand on this definition by saying that it is also describable.

An *organization* is a specific type of resource that brings people together to participate in VO applications. Organizations can be hierarchical and range greatly in size and scope. At a high level, an organization could be a university, observatory, or government agency. At a finer level, it could be a specific scientific project, space mission, or individual researcher. A provider is an organization that makes data and/or services available to users over the network.

A *service* is any VO resource that can be invoked by a user or software agent to perform some action on their behalf. Associated with any service is descriptive metadata about the service. This metadata generally include information the user needs to determine if a service is of interest and how the service may be invoked. Specific types of metadata are described below. Note that the service itself need not be aware of the metadata that describe it.

A *query service* supports a query/response protocol. The user submits a query to the service that may define characteristics of interest, and the service returns a set of information to the user. The query may be null, e.g., a current-time service may only support a null query, and some services may respond to a null query with appropriate default actions. Non-query services may also exist, e.g., services to copy or delete files on remote file systems, to mail information to other users, to kill existing jobs, to authorize actions, etc.

A *registry* is a service which aggregates and serves resource metadata. The metadata may be added to the registry via an input form or harvested from the resources themselves. A registry may serve all resource metadata (full registry), select types of resource (limited registry) or resources at a specific location (local registry). Any registry may also support a query interface which will allow searching for resources based on various combinations of metadata values.

A sample of the metadata that would be used to describe the Sloan Digital Sky Survey source catalog as hosted at the Space Telescope Science Institute is shown in Fig. 1. Further information concerning the encoding of such metadata and their incorporation into resource registries is describe by Plante et al. (2004) and Greene et al. (2004).

2. Lessons Learned and Questions Raised in Populating a Prototype Registry

Both the NVO and AstroGrid projects have implemented prototype registries. The NVO prototype has been used as a data discovery engine for the Data Inventory Service (http://heasarc.gsfc.nasa.gov/vo/data-inventory.html, McGlynn et al. 2004). The prototype registry was constructed primarily through manual entry of metadata about known cone search and Simple Image Access

Identity metadata	
Title	Sloan Digital Sky Survey
ShortName	SDSS
Identifier	ivo://stsci.edu/mast/sdss

Curation metadata	
Publisher	Space Telescope Science Institute/MAST
PublisherID	ivo://stsci.edu/mast
Creator	Sloan Digital Sky Survey Consortium
Creator.Logo	http://archive.stsci.edu/images/sdss_logo.gif
Contributor	Sloan Digital Sky Survey Consortium
Date	2001-06-15
Version	SDSS EDR
ReferenceURL	http://archive.stsci.edu/sdss/index.html
Contact.Name	Archive Branch, Space Telescope Science Institute
Contact.Address	3700 San Martin Drive, Baltimore, MD 21218 USA
Contact.Email	archive@stsci.edu
Contact.Telephone	+1-410-338-4547

General content metadata	
Subject	galaxies, quasars, stars, CCD photometry, spectroscopy, redshift, sky surveys
Description	The Sloan Digital Sky Survey is using a dedicated 2.5-m telescope and a large format CCD camera to obtain images of over 10,000 square degrees of high Galactic latitude sky in five broad bands (u', g', r', i' and z', centered at 3540, 4770, 6230, 7630, and 9130 A, respectively). . . .
Source	2002AJ...123..485S
Type	Survey, Catalog, EPOResource
ContentLevel	Research
Relationship	mirror-of
RelationshipID	ivo://sdss.org/sdss/edr

Collection and service content metadata	
Facility	Apache Point Observatory, Sloan 2.5-m Telescope
Instrument	Five-band clocked CCD camera
Coverage.Spatial	polygon (FK5, 145.17, 1.25, 235.9, 1.25, 235.9, -1.25, 145.17, 1.25) or polygon (FK5, 250.71, 66.29, 267.0, 66.29, 267.0, 52.15, 250.71, 66.29) or polygon (FK5, 350.43, 1.17, 360.0, 1.17,360.0, -1.25, 350.43, -1.25) or polygon (FK5, 0.0, 1.17, 56.37, 1.17, 56.37, -1.25, 0.0, -1.25)
Coverage.RegionOfRegard	0.0001
Coverage.Spectral	Optical
Coverage.Spectral.Bandpass	u, g, r, i, z
Coverage.Spectral.MinimumWavelength	400.e-9
Coverage.Spectral.MaximumWavelength	850.e-9
Coverage.Temporal.StartTime	1999-12-25
Coverage.Temporal.StopTime	2001-07-15
Coverage.Depth	3.e-6
Coverage.ObjectDensity	6.e4
Coverage.ObjectCount	2.e7
Coverage.SkyFraction	0.01
Resolution.Spatial	0.00028
Resolution.Spectral	5000
Resolution.Temporal	120
UCD	Not Provided
Format	text/xml
Rights	Public

Data quality metadata	
DataQuality	A
Uncertainty.Photometric	3.e-7
Uncertainty.Spatial	0.00003
Uncertainty.Spectral	1.e-11
Uncertainty.Temporal	0.1

Service metadata	
Service.InterfaceURL	http://archive.stsci.edu/cgi-bin/sdss/catalog.html
Service.BaseURL	http://archive.stsci.edu/cgi-bin/sdss/catalog
Service.HTTPResults	text/xml
Service.StandardID	ivo://ivoa.net/Services/ConeSearch
Service.StandardURL	ivo://www.ivoa.net/Documents/REC/ConeSearch.html
Service.MaxSearchRadius	0.2
Service.MaxReturnRecords	5000

Figure 1. Sample resource metadata. Dublin Core elements are shown in bold, and required elements are shown in italics. (Bold italics indicate required elements that are also in the Dublin Core.) See http://dublincore.org for more information about Dublin Core metadata.

Protocol services. It took about a week to populate a prototype registry of ∼100 resources. During this time period, we recognized certain patterns in data entry as well as inconsistencies in metadata descriptions. This experience leads to the following conclusions and questions:

- Don't ask for too much metadata: publishers will not enter, or will enter inaccurate information.
- Provide guidance in metadata entry and interpretation. Definitions must be clear and unambiguous. Be inclusive.
- Metadata entry should be as automated as possible. Need interactive entry tools with pull-down pick-lists, for example.
- Standardize units. Interfaces can perform conversions if necessary.
- The syntax and semantics for Identifier need to be finalized, and experience gained in just how Identifiers will be used.
- How should the spatial (angular) resolution of a resource be characterized? By "best" or "worst"? ("Best" is most consistent with being inclusive.)
- How specific/complex should spatial, spectral, and temporal coverage be?
- Need agreement on how to specify "not known", "not applicable", and "not provided", including for numeric values.
- Should all metadata elements be explicitly typed?

In addition, the resource metadata concepts described here must be encoded and structured in a machine-readable registries. Work continues on XML schema that more fully show the relationships among metadata elements and that simplify data entry and maintenance efforts (e.g., by allowing an organization to register its curation-related metadata once and apply it to a number of different collections).

References

Berners-Lee, T., Fielding, R., & Masinter, L. 1998, IETF RFC2396, http://asg.web.cmu.edu/rfc/rfc2396.html

Greene, G., O'Mullane, W., Hanisch, R., & Gaffney, N. 2004, this volume, 285

McGlynn, T., Lee, J., Hanisch, R., O'Mullane, W., & Greene, G. 2004, this volume, 319

Plante, R.,Greene, G., Hanisch, R., McGlynn, T., O'Mullane, W., Williams, R., & Williamson, R. 2004, this volume, 585