

Resource and Service Metadata for the Virtual Observatory
--- DRAFT, 7 June 2002 ---

NVO Metadata Working Group

Introduction

An essential capability of the Virtual Observatory is a means for describing what data and computational facilities are available where, and once identified, how to use them. The data themselves have associated metadata (e.g., FITS keywords), and similarly we require metadata about data collections and data services so that VO users can easily find information of interest. Furthermore, such metadata is needed in order to manage distributed queries efficiently; if a user is interested in finding x-ray images there is no point in querying the HST archive, for example. In this document we suggest an architecture for resource and service metadata and describe the relationship of this architecture to emerging Web Services standards. We also define an initial set of metadata concepts.

Architecture

In order to make it easy for astronomy information services to participate in the VO, we propose a hierarchical system for metadata management. 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 is more extensive and complex, allowing for the description of query syntax, access protocols, and usage policies.

A *service* is any VO element that can be invoked by the user to perform some action on their behalf. Associated with any service is descriptive *metadata* about the service. Metadata generally includes 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 describes 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 files systems, to mail information to other users, to kill existing jobs, to authorize actions, etc.

A *registry* is a query service for which the response is a structured description of other services. The services described by a registry may be of any type. The registry may support a query that allows the user to indicate which services might be of interest.

A *resource* is a collection of one or more services, or other resources, that share some common metadata characteristics (e.g., Publisher, Creator, Contributor, Identifier, Contact, Type, Facility). The extent of commonality depends upon the resource. The services described by a resource may themselves be resources. For example, MAST, HEASARC, IRSA, NED, et al., are resources. Each of these contains other resources, e.g., the HST archive in MAST. They also contain specific services, such as an HST observation log query service or a cone search service. A resource must include at least a minimalist service, i.e., a URL for a web site. One could in principle describe all of NASA astrophysics data holdings as a resource, or all of NVO as a resource, but aggregates of this scale circumvent the goal of being able to locate the specific resources and services of interest for a particular application.

Both resources and services are described by metadata. *Resource metadata* are high-level and independent of any specific service. Resource metadata include

- *Curation metadata*, which describe who supports the resource and what its purpose is
- *Content metadata*, which describe what kind of information is available (types of data, sky coverage, spectral coverage, etc.)

Resource metadata are typically not queryable parameters in the underlying services, but rather they encompass information that now is simply “known” to users, or must be discovered through other means. Astronomers know that the HST archive includes optical images and spectra, for example, or that Vizier provides access to catalogs and tables. Resource metadata constitute a “yellow pages” of astronomical information. Resource metadata are analogous to the UDDI (Universal Description, Discovery and Integration) Web Service, and are analogous to the high-level descriptions included in the CDS GLU.

Service metadata include metadata that describe the service's interface (its input and output) as well as information that aids in effective use of the service (e.g., range of possible values returned). Service metadata also describe access methods or protocols. Service metadata are analogous to WSDL (Web Service Description Language) and the query specifications component of the CDS GLU.

These analogies are not perfect. For example, WSDL can describe multiple services in one file, and UDDI does probably not convey as much information as resource metadata. Nevertheless, the intention is for the resource metadata to describe *what* is available, and for the service metadata to describe *how* to access it.

Resource Metadata Concepts*

Below we describe the *concepts* we believe are needed in the resource metadata. The names of these concepts need to be encoded in a standard representation, such as UCDs.

Curation Metadata

Title (string)

Definition: A name given to the resource.

Comment: Typically, a Title will be a name by which the resource is formally known.

Publisher (string)

Definition: An entity responsible for making the resource available

Comment: Examples of a Publisher include a person or an organization.

Creator (string)

Definition: An entity primarily responsible for making the content of the resource.

Comment: Examples of a Creator include a person or an organization.

Description (string, free text)

Definition: An account of the content of the resource.

Comment: Description may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content.

Contributor (string)

Definition: An entity responsible for making contributions to the content of the resource.

Comment: Examples of a Contributor include a person or an organization.

* Resource metadata concepts are drawn from the Dublin Core, <http://dublincore.org/documents/dces/>, except where otherwise noted.)

Identifier (URI)

Definition: An unambiguous reference to the resource within a given context.

Comment: The URI corresponding to the resource.

Reference URL (URL)

Definition: A URL pointing to additional information about the resource.

Comment: Not in Dublin Core.

Contact (string, e-mail address)

Definition: The e-mail address for contacting the persons responsible for the resource.

Comment: Not part of the Dublin Core. *Contact* is split into two concepts for clarity.

Contact.Name (string)

Definition: The name of the contact.

Comment: A person's name, "John P. Jones", or a group, "Archive Support Team".

Contact.Email (e-mail address)

Definition: The e-mail address of the contact.

Comment: For example, "mailto:John.P.Jones@navy.gov", or "mailto:archive@datacenter.org".

Content Metadata

Type (string, list)

Definition: The nature or genre of the content of the resource.

Comment: Type includes terms describing general categories, functions, genres, or aggregation levels for content. Recommended best practice is to select a value from a controlled vocabulary. VO *Types* include, e.g., archive, survey, catalog, bibliography, journal, library. Resources providing more than one type of content would simply list all relevant types.

Coverage (string)

Definition: The extent of scope of the content of the resource.

Comment: The Dublin Core notion of coverage is too generic to be of much use in the VO, where we need more specific information. We propose to subset this element as follows:

Coverage.Spatial (string)

Definition: The sky coverage of the resource.

Comment: This will require an agreed syntax.

Coverage.Spectral (string, list)

Definition: The spectral coverage of the resource.

Comment: Spectral coverage at the resource level will be in terms of general bandpasses (gamma-ray, x-ray, extreme UV, UV, optical, infrared, microwave, radio). Multiple bandpasses can be given for high-level resources. We will require use of a controlled vocabulary, and the bandpasses will need to be defined specifically (e.g., "optical" is $3200 \text{ \AA} \leq \lambda \leq 8000 \text{ \AA}$).

Coverage.Temporal (string)

Definition: The temporal coverage of the resource.

Comment: This will require an agreed syntax.

Facility (string)

Definition: The observatory or facility where the data was obtained.

Comments: Not in Dublin Core. Some resources are likely to hold data from multiple observatories. If just a few, this could be a list; if very many, just say “many”. Theoretical data will not originate with an observatory, thus the more generic content label of “facility”.

Instrument (string)

Definition: The instrument used to collect the data.

Comments: Not in Dublin Core. Can be a specific instrument name (Wide Field/Planetary Camera 2) or generic instrument type (CCD camera). Theoretical data is produced by a computer code, and the name of the code could be specified.

Format (string)

Definition: The encoding format of data provided by the resource.

Comments: Typical values would be “FITS”, “ASCII text”, “HTML”, “XML”, “VOTable”, “GIF”, etc. Dublin Core notion of Format is different, but very flexible. Recommend employing MIME types here in order to utilize existing standards.

Rights (string)

Definition: Information about rights held in and over the resource.

Comment: Dublin Core uses Rights to describe copyright and other intellectual property rights issues. In the VO context Rights would describe access privileges (public, proprietary, mixed, with a defined vocabulary).

Resource metadata would typically be collected through a resource registration service, i.e., a web form that would present a resource curator with the requisite fields and enumerated lists, and would construct a resource descriptor in a standard format (such as VOTable). If content elements are not relevant for a given resource Type (e.g., Type “journal” does not have meaningful spatial coverage, Facility, or Instrument, though it does have a valid temporal coverage), then they should take on a “not applicable” value. The resource registration service should not allow fields to be left unspecified.

Example: The Sloan Digital Sky Survey data as hosted by MAST at STScI.

| | |
|-------------------------|--|
| <i>Title</i> | Sloan Digital Sky Survey |
| <i>Publisher</i> | Space Telescope Science Institute/MAST |
| <i>Creator</i> | Sloan Digital Sky Survey Consortium |
| <i>Description</i> | 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 Å, respectively). Medium resolution spectra will be obtained for approximately 10 ⁶ galaxies and 100,000 quasars. The early data release (EDR), on June 2001, includes searchable catalogs of images and spectra, images for display and scientific purpose in both 2-D FITS and JPEG formats, and spectra in both 1-D FITS and GIF formats. The EDR covers about 460 square degrees of sky. The next data releases will occur every 18 months or so. |
| <i>Contributor</i> | Sloan Digital Sky Survey Consortium |
| <i>Identifier</i> | http://archive.stsci.edu/sdss/ |
| <i>ReferenceURL</i> | http://archive.stsci.edu/sdss/index.html |
| <i>Contact.Name</i> | Archive Branch, Space Telescope Science Institute |
| <i>Contact.Email</i> | mailto:archive@stsci.edu |
| <i>Type</i> | survey, catalog |
| <i>Coverage.Spatial</i> | (RECT J2000 145.17 -1.25 235.9 1.25) OR (RECT J2000 250.71 52.15 267 66.29) OR (RECT J2000 350.43 -1.25 416.37 1.17) |

| | |
|--------------------------|--|
| <i>Coverage.Spectral</i> | optical, infrared |
| <i>Coverage.Temporal</i> | 1999- |
| <i>Facility</i> | Apache Point Observatory, Sloan 2.5-m Telescope |
| <i>Instrument</i> | Five-band clocked CCD camera |
| <i>Format</i> | FITS, GIF, JPEG (image/FITS, image/gif, image/jpg) |
| <i>Rights</i> | public |

Service Metadata Concepts

[These ideas are only partially fleshed out.]

Since we have defined a *resource* as containing one or more *services* (or one or more *resources*), we wish to show the relationship between resources and the services they contain, and between services and the resources they belong to. We have also said that a resource can contain multiple services that share resource metadata. Service metadata therefore inherit resource metadata from the resource in which they are contained. Service metadata may expand upon resource metadata, adding specificity, or override resource metadata in cases where it is impossible to define resource metadata explicitly. For example, a resource such as Vizier covers, in principle, the entire sky (*Coverage.Spatial*), but specific catalog services within Vizier will have more limited coverage.

Discussion has only just begun as to whether or not the resource metadata concepts should explicitly include a *Service* or *Services* element.