VOResource and DataCite: Converging Metadata Schemes

Markus Demleitner[1,⋆]

[1] *Universität Heidelberg, Zentrum für Astronomie, Astronomisches Rechen-Institut*

Abstract. For the past 10 years, VOResource, together with its extensions, has been the metadata schema used for service description and discovery in the Virtual Observatory (VO). We are currently evolving the core VOResource schema into a version 1.1, in particular to ensure smooth interoperability with the DataCite metadata kernel. In this contribution, we will give a brief overview of how and why VOResource has supplemented Dublin Core with "actionable" metadata, as well as how the current evolution will update it to provide easy paths for VO data providers into the DataCite world.

1 Introduction

The Virtual Observatory (VO) is a large network of astronomy-related data services and data collections, held together by commonly implemented protocols and a directory of services available. This directory is called the VO Registry and is maintained by an informal association of individuals from various institutions from around the world known as the Registry Working Group (RWG) of the IVOA (International Virtual Observatory Association).

The VO Registry conceptually is the collection of metadata records (currently roughly 17000). In the next section, we briefly illustrate the most common uses for this metadata collection.

The technical realisation of the VO Registry borrows heavily from digital library technology. In the third section, we discuss how the use cases motivate the additions the VO community has made to the basic technologies.

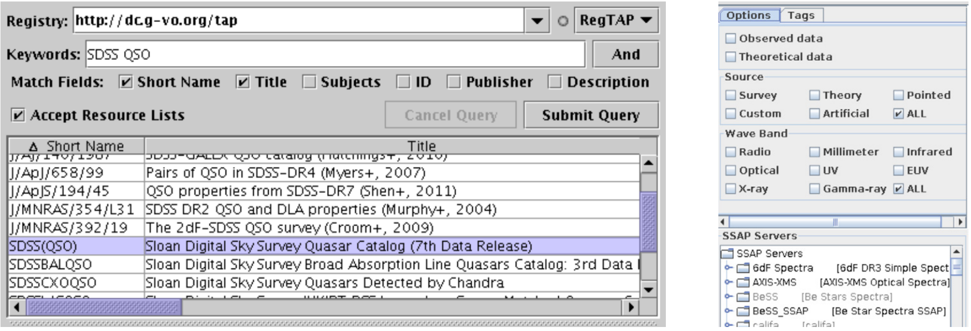
The RWG has a strong interest in maintaining interoperability with library systems and data directories outside of astronomy, in particular regarding establishing good practices of data citation. We are therefore currently adapting our main metadata schema to facilitate workflows involving both DataCite and VO components. The fourth section discusses these recent initiatives.

We conclude with considerations on what further measures can be taken to improve interoperability between VOResource and DataCite.

2 Common VO Registry Use Cases

Currently, most VO users encounter the Registry through in-application interfaces. For instance, the popular application TOPCAT [1] lets users select services to query by matching keywords against a user-selectable set of natural language fields; the actual Registry query is hidden from the user (Fig. 1 a). Another application, SPLAT [2], uses the Registry to discover all spectral services and

[⋆] e-mail: msdemlei@ari.uni-heidelberg.de



Registry: http://dc.g-vo.org/tap ○ RegTAP

Keywords: SDSS QSO

Match Fields: ☑ Short Name ☑ Title □ Subjects □ ID □ Publisher □ Description

☑ Accept Resource Lists

Short Name	Title
J/ApJ/658/99	Pairs of QSO in SDSS-DR4 (Myers+, 2007)
J/ApJS/194/45	QSO properties from SDSS-DR7 (Shen+, 2011)
J/MNRAS/354/L31	SDSS DR2 QSO and DLA properties (Murphy+, 2004)
J/MNRAS/392/19	The 2dF-SDSS QSO survey (Croom+, 2009)
SDSS(QSO)	Sloan Digital Sky Survey Quasar Catalog (7th Data Release)
SDSSBALQSO	Sloan Digital Sky Survey Broad Absorption Line Quasars Catalog: 3rd Data
SDSSCXOQSO	Sloan Digital Sky Survey Quasars Detected by Chandra

□ Observed data
□ Theoretical data

Source
□ Survey □ Theory □ Pointed
□ Custom □ Artificial ☑ ALL

Wave Band
□ Radio □ Millimeter □ Infrared
□ Optical □ UV □ EUV
□ X-ray □ Gamma-ray ☑ ALL

SSAP Servers

SSAP Servers
6dF Spectra [6dF DR3 Simple Spect
AXIS-XMS [AXIS-XMS Optical Spectra]
BeSS [Be Stars Spectra]
BeSS_SSAP [Be Star Spectra SSAP]

Figure 1. Examples for in-application registry interfaces. (a) TOPCAT, left; (b) SPLAT, right.

defaults to querying a subset of these services with one user interaction. This subset is selectable by defining categories of interest (Fig. 1 b). Again, end-users are shielded from the actual operation of the Registry.

The upcoming version 10 of the Aladin image processing application (cf. [3]) will feature a very interactive integrated resource discovery mode where, while scrolling over the sky, a list of services of interest to the user is updated by coverage and criteria like spectral region or the presence of certain kinds of data (proper motions, say) in the services' data schemas.

There are more conventional OPAC-like interfaces, for instance the Web Interface to the Relational Registry WIRR[1] or analogous services operated by ESA[2] or the STScI[3]. They are, however, used much less in comparison to in-application facilities.

The VO Registry has further uses in infrastructure maintenance. For instance, validators use Registry information to discover service endpoints to operate on, and to determine what queries will yield validatable results[4].

For a relatively recent review of Registry client interfaces, see [4].

3 Adaptation of Library Technology

The VO Registry is designed as a fairly conventional distributed system, where publishing registries are harvested by metadata aggregators using OAI-PMH. The relatively few extra rules are laid down in a standard called *Registry Interfaces* [5].

The central components typically are searchable registries. The normal browser-based OPACs one would typically see for libraries of texts are not really sufficient for the VO. For one, searchable registries must interoperate with the clients mentioned in sect. 2, which requires a well-defined interface in terms of both query parameters and the response schema. Also, the VO's complex metadata schema does not lend itself to simple web displays.

Therefore, in the VO the standard searchable registries support a non-library protocol called RegTAP [6], which is essentially a combination of a database schema (13 tables as of RegTAP 1.0) and the VO Table Access Protocol TAP [7], which defines how to transport database queries and their results. This obviously does not preclude the provision of searchable registries supporting non-standard interfaces, as illustrated by the web interfaces mentioned above.

[1] http://http://dc.zah.uni-heidelberg.de/wirr/q/ui/fixed
[2] http://registry.euro-vo.org/eurovo/
[3] http://nvo.stsci.edu/vor10/index.aspx
[4] e.g., https://heasarc.gsfc.nasa.gov/vo/validation/vresults.pl

As required by OAI-PMH, publishing registries in the VO can deliver their records in the OAI Dublin Core metadata schema. Service discovery, however, very typically involves queries like "Give me the access URLs of all services speaking version 1 of an image access protocol with infrared data" or "Give my database services with tables about quasars having redshifts and radio fluxes". To support such use cases, the VO Registry has the *ivo_vor* metadata schema, specified in the VOResource standard [8] and a set of extensions.

Features beyond Dublin Core include, in particular:

- capabilities – these define modes of data access, i.e., support for certain protocols; the protocols are defined using identifiers into the Registry itself. For instance, a service could say that its underlying data collection is available through a standard image search as well as through the Table Access Protocol mentioned above. In the latter capability the service could furthermore declare that the published table conforms to a pre-defined table schema ("data model"). This is actually how searchable registries are currently discovered, and is used in making collections of observational data products uniformly accessible with advanced query modes, including matches against bulk data uploaded by the client.
- interfaces – these link the capabilities to access URLs ("endpoints"). Additional metadata, such as names, types, and units for parameters supported on such an interface or the type of data returned, can be given. Interface metadata can also contain information on supported authentication methods, although only very few services actually give such metadata in current practice.
- tablesets – these describe the table structure of the data underlying a service, including name, type, unit, and physical characterisation of the columns that make up the tables. Tablesets enable the "discovery by physics" use case hinted at above.
- coverage in space, time, and spectrum – while one might suspect that this is very basic metadata for an astronomical data service, VO practice only now starts to properly support discovery by coverage. The reason for the sluggish definition of physical coverage in VO resource records is probably technological. Only the emergence of MOCs [9] – a compact, healpix-based representation of arbitrary shapes on a sphere – has made the definition and processing of spatial coverages straightforward enough for widespread adoption.

4 Convergence with DataCite

In particular with a view to facilitate consistent citation of data sources used in scholarly publications and the provenance of derived data, an overhaul of VOResource 1.0 was started. The focus was on enabling smooth interoperability between VOResource and the DataCite metadata kernel [10], which forms the basis of DOI registration. We expect to finish this update (VOResource 1.1) in 2017. It is currently under public review, the results of which are preserved in the IVOA's discussion system[5].

Major additions and changes in VOResource 1.1 with respect to DataCite interoperability include:

- records, creators, and contacts now support zero or more *altIdentifier*s. These will in general contain DOIs for records, and ORCIDs for persons, but the identifier type is intended to be inferred by the identifiers' URI schemes. The VO's own identifier scheme, *ivo:*, remains unchanged.
- VOResource had terms for *relationship* types that are incompatible with DataCite both in style (dash-joined instead of CamelCase) and content (e.g., DataCite has no served-by, VOResource's mirror-of roughly matches DataCite's IsIdenticalTo). We have taken over DataCite's terms but cannot entirely drop the old terms for backward compatibility. An RDF document now describes

[5] http://wiki.ivoa.net/twiki/bin/view/IVOA/VOResource11RFC

the relations between the terms and adds several terms not present in DataCite's vocabulary but necessary for VO operation (in particular isServiceFor and IsServedBy).

- Similarly, the terms for date roles were harmonised. This was mainly a matter of mapping word types, for instance VOResource's "creation" to DataCite's "Created".
- DataCite's model of *rights* and *rightsURI* was simply copied; VOResource 1.0's *rights* element was so restricted that it has not been used at all and thus could be replaced without adverse consequences.

As a result of these changes, an XSLT style sheet[6] can translate VOResource metadata to DataCite records almost ready for DOI minting. Additional logic is merely required to actually form the DOIs themselves.

5 Further Work

For more complete preservation of the contents of VOResource records in DataCite metadata, it would be desirable to

1. Add the extra terms *IsServiceFor* and *IsServedBy* to DataCite's *relationType* vocabulary.
2. Add the term *IVOID* to DataCite's *relatedIdentifierType* vocabulary. This is not strictly necessary, as in principle *URL* could be used with the nature of the identifier being inferred by the URI scheme. However, since the *ivo:* scheme is uncommon outside of the Virtual Observatory, an additional hint as to how the identifiers are to be resolved appears desirable.

At least for the forseeable future, an inclusion of extra VOResource metadata (capabilities, tablesets) into the DataCite metadata kernel is probably neither feasible – as seen by the size of the Registry XML schema documents compared to DataCite alone – nor desirable – as it seems unlikely generic DataCite metadata consumers will interpret the metadata any time soon.

On the VO side, the changes in VOResource 1.1 have to be translated into RegTAP's database schema. Work on this, in the form of an early working draft for RegTAP 1.1, is underway.

References

[1] M.B. Taylor, *TOPCAT & STIL: Starlink Table/VOTable Processing Software*, in *Astronomical Data Analysis Software and Systems XIV*, edited by P. Shopbell, M. Britton, R. Ebert (2005), Vol. 347 of *Astronomical Society of the Pacific Conference Series*, p. 29

[2] M. Castro-Neves, P.W. Draper, *SPLAT-VO: Spectral Analysis Tool for the Virtual Observatory* (2014), astrophysics Source Code Library, `1402.008`

[3] F. Bonnarel, P. Fernique, O. Bienaymé, D. Egret, F. Genova, M. Louys, F. Ochsenbein, M. Wenger, J.G. Bartlett, Astronomy and Astrophysics Supplement **143**, 33 (2000)

[4] M. Demleitner, P. Harrison, M. Taylor, J. Normand, Astronomy and Computing **11**, 91 (2015)

[5] K. Benson, R. Plante, E. Auden, M. Graham, G. Greene, M. Hill, T. Linde, D. Morris, W. O'Mullane, G. Rixon et al., *IVOA Registry Interfaces Version 1.0*, IVOA Recommendation 04 November 2009 (2009), `1110.0513`

[6] M. Demleitner, P. Harrison, M. Molinaro, G. Greene, T. Dower, M. Perdikeas, *IVOA Registry Relational Schema Version 1.0*, IVOA Recommendation 08 December 2014 (2014), `1510.02275`

[6] https://volute.g-vo.org/svn/trunk/projects/registry/dois/vor-to-doi.xslt

[7] P. Dowler, G. Rixon, D. Tody, *Table Access Protocol Version 1.0*, IVOA Recommendation 27 March 2010 (2010), `1110.0497`

[8] R. Plante, K. Benson, M. Graham, G. Greene, P. Harrison, G. Lemson, T. Linde, G. Rixon, A. Stébé, IVOA Registry Working Group, *VOResource: an XML Encoding Schema for Resource Metadata Version 1.03*, IVOA Recommendation 22 February 2008 (2008), `1110.0515`

[9] P. Fernique, T. Boch, T. Donaldson, D. Durand, W. O'Mullane, M. Reinecke, M. Taylor, *MOC - HEALPix Multi-Order Coverage map Version 1.0*, IVOA Recommendation 02 June 2014 (2014), `1505.02937`

[10] DataCite Metadata Working Group, *DataCite metadata schema – documentation for the publication and citation of research data version 4.0*, DataCite publication (2016), `https://schema.datacite.org/meta/kernel-4.0/doc/DataCite-MetadataKernel_v4.0.pdf`