

DCP Notes

Wednesday 4 June 2025

Presentation: Chenzhou Cui, China VO

Metadata for data sets is maintained by the National Astronomical Data Center, including provenance data. But this is not meeting China-VO needs for discovery and data service. A multi-dimensional metadata framework specification is in development. The intention is to design in interoperability with IVOA, DOI, CSTR, and other metadata contexts. Six metadata schema standards were analyzed and compared for commonalities and to produce cross-walks between them. (Slides contain a comparison table of various attributes and their presence/absence in various contexts.) Merging attributes across the publishing contexts (Dublin Core, DOI, IVOA) and the Chinese context, and adding additional fields specific to China-VO yielded 34 required and 57 optional fields.

The list of data types produced is more specific than IVOA has in its vocabulary at the present time (less than twice as long). Additional classifications are included in the presentation slides.

Also in the archive are "Paper Data" datasets - datasets specifically related to a publication. The naming of these data sets is somewhat arbitrary - sometimes it is the name of the paper, sometimes it describes the data, sometimes something in between

Gus notes that, in terms of conventions, the title should certainly describe the data, and it should certainly *not* be the title of the ~~data~~ article. Chenzhou notes this is beyond the control of the IVOA.

Gilles notes that link by coincident title is not a reasonable way to associate copies of the data that reside in different places.

Chenzhou asks if there is a way to deposit Chinese paper data into Vizier. Gilles responds that it is not - that it begins with the paper. If the paper prompts interest in the data, then the data will be sought for inclusion. There may be scope for improving the path for China-VO.

Markus asks why you would want to put data in Vizier that is already in China-VO. Chenzhou notes it is a matter of order and repository chosen. The data typically comes as an adjunct to a paper, and repositories collecting paper data (like Zenodo, for example) do not provide services for curating or discovering the data.

Chenzhou asks if it is possible to create an ADS link for the paper data.

Presentation: Gilles Landais, Data Origins in VO

Extracting data origin information with a new AstroPy module, from a VOTable. Also a demonstration of working with provenance information. Demonstrated live via python notebook. Included working with the provenance graph.

Presentation: Pierre Le Sidenar; Data Cite and IVOA object descriptions

DataCite tends to provide finer granularity and precision for creators, affiliations, contributors.

Gus notes that the concept of “contributor” adds some confusion, in that it is not clear how/when/if single people who make multiple contributions (creator, sponsor, etc.) should be listed multiple times.

In addition, DataCite provides a number of relationship types for connecting resources that are somehow related. These include “isNewVersionOf” and “IsPreviousVersionOf”. Pierre proposes that these should be used in VOResource when the relationships exist.

The resource types used by DataCite are many and varied, but for the most part are not related to the kind of VO resource types relevant to VO queries. VOID is also not persistent in the way that DOIs are. Persistence is a key factor for DOIs.

SPASE metadata was then compared to both DataCite and IVOA.

Gus notes the VO Resource type appears to be domain-specific.

Stephane asks about creators: There is an ambiguity between who set up the (VO) service, and who provided the data. Crediting the party who created the resource but not the data creator can cause friction between the data and service providers. Baptiste suggests this is related to the question of what the correct title of the resource should be. Stephane argues that, for example, a PI will never be happy when a service provider is acknowledged and the PI is not.

Markus notes that most people seem to not understand the details of the fairly (possibly too) rich data model. Should PIs just get used to not being constantly acknowledged.

Gus asks whether this is a metadata problem or an ethical problem, or possibly both.

There does not appear to be disagreement on what the “creator of a resource” is, but when a service is set up, it is not necessarily clear to the service creator what the “resource” is - the data or the service.

Presentation: Tim Jeness; DOI Plans for Rubin

Data Preview 1 scheduled for release June 30, and will include DOIs. The expectation is that DOIs will be used to connect various parts of the data and of the system, including instrumentation. Facilities apparently cannot have DOIs.

Review of “DOI Proposal” Slide.

Anne R. notes that “IsIdenticalTo” has been problematic for DataCite providers, and should probably never be used outside of the case where two DOIs are assigned to precisely the same incarnation of a resource and resolve to the same place. Everything else tends to call under “IsVariantFormOf”

Brief discussion about the use of DOIs, vs PIDs (permanent Identifiers) from other sources. DataCite provides generic metadata for many types of things and also some services to make linkages to other DataCite and partner (like CrossRef) metadata providers.

To the Best Practices Document...

Summary of scope and present top-level outline.

Big Questions:

- What is the right level of metadata?
- What role should VOResource play?
- What are the implications of assigning DOIs to parts or aggregates of data products, to instruments, etc.?

The interpretation of metadata fields, especially in regards to creating references, is problematic. Questions are constantly arising as to what the metadata appearing in a reference means (what does the “publication year” mean in the context of an instrument, for example).

Tim J. notes that the DataCite input form for creating a DOI is more than a little complicated and difficult to map to, for example, and instrument context. Anne R. suggests mapping the schema fields from the DataCite documentation to the specific domain first, so that at least there is some consistency in interpretation.