Astronomy and Computing 11 (2015) 74–80

Contents lists available at ScienceDirect

Astronomy and Computing

journal homepage: www.elsevier.com/locate/ascom

Full length article

# Data modeling for the virtual observatory

Mireille Louys *

*Centre de Données astronomiques de Strasbourg, Observatoire astronomique de Strasbourg, Université de Strasbourg, CNRS, UMR 7550, 11 rue de l' Université, F-67000 Strasbourg, France*
*ICube, Université de Strasbourg, CNRS UMR 7357, 300 bd Sébastien Brant - F-67412 Illkirch Cedex, France*

ARTICLE INFO

*Article history:*
Received 3 November 2014
Received in revised form
6 March 2015
Accepted 6 March 2015
Available online 18 March 2015

*Keywords:*
Data modeling
Astronomical databases
Virtual observatory tools

ABSTRACT

The data modeling effort has played a key role in the Virtual Observatory project, and contributed to the effort to build a common reference framework to describe the necessary information attached to astronomical data: the metadata. Such metadata describe the observing parameters and characterize and qualify the observed measurements. These pieces of information are produced and stored in project archives. Standardizing a homogeneous representation of metadata allows uniform discovery and use of the data in the Virtual Observatory infrastructure. This paper describes the context of data modeling in the VO architecture and shows how data models support requirements on the data access layer and for applications development. How the modeling process has been undertaken is explained with a short overview of the different data models. We also discuss in some detail the lessons learned in this modeling and standardization effort.



## 1. Introduction

This paper highlights the collaborative work undertaken in the Virtual Observatory (VO) project for modeling the observational metadata published by various astronomical data centers and used by scientists for their research programs. We present an overview of how the data modeling effort in the Data Model Working Group (DM WG), gathered and structured knowledge about observations and their metadata descriptions in a set of articulated data models.

During the VO development a number of other modeling efforts have been undertaken. The VOEvent group defined a description of sky events, with an adhoc protocol and data model (Seaman et al., 2011). Simulation codes and simulation data are modeled in a specific top-down approach (Lemson et al., 2014) led by the Theory interest group and endorsed by the DM WG. In this paper we concentrate on observational data, and explain the DM WG effort and its interactions with archive data providers, the Data Access Layer Working Group, and the applications developers. This work has involved strong interaction with other working groups efforts and developments. Section 2 outlines the approach to observational data. Section 3 describes the IVOA landscape and Section 4 provides details on the data modeling process. The current data models are described in Section 5 with discussion of the lessons learnt in Section 6. The *Glossary* section at the end of the paper defines the acronyms and development tools currently used in the VO initiative.

## 2. A dedicated approach for observational data

The VO data modeling effort is intended to organize and offer a description of the observational datasets in a logical and comprehensive way. It encodes common reference knowledge about metadata associated with observations that helps users navigate on-line distributed data collections, and allows publishers to efficiently describe their resources. It is also driven by science cases and the necessity to sort out, compare and confront data files from different observation programs. Interoperability has become a common concern for most disciplines in observational sciences; this has led to the emergence of similar projects such as Helio-VO[1] in heliophysics, or VAMDC[2] for atomic and molecular physics.

In the context of distributed astronomical science products, data are generally public after some proprietary period. The scientific value of data lasts for a long time. There is intrinsic value in data from different epochs for understanding time-dependent

* Correspondence to: Centre de Données astronomiques de Strasbourg, Observatoire astronomique de Strasbourg, Université de Strasbourg, CNRS, UMR 7550, 11 rue de l' Université, F-67000 Strasbourg, France.
*E-mail address:* mireille.louys@unistra.fr.

http://dx.doi.org/10.1016/j.ascom.2015.03.004
2213-1337/

[1] http://www.helio-vo.eu/.
[2] http://www.vamdc.eu/.

phenomena. The diversity of observing programs provides rich, heterogeneous data in many forms: light curves, spectra, spectral energy distributions, sky images, and velocity or spectral datacubes.

Astronomy data are stored for long term preservation so that they can be re-used in various studies later on. It is important to trace the data precision and statistical signature of each dataset individually, in order to characterize the content. Multi-wavelength studies need precise descriptions of the instrumental parameters and of the statistical behavior of the measurements in order to reliably compare, match, superimpose, or combine observations in various regimes.

Moreover astronomy archives use a wide variety of database systems and architectures, with each organization using its own design for table definition, column names, etc. Therefore in order to allow scientists to seamlessly access a large collection of various archives, without having to learn each particular interface for accessing data of interest, there was a need for a common description frame for all astronomical metadata. This has been a strong incentive for the development of data modeling in the Virtual Observatory project and to build up protocols and applications in a consistent manner.

The approach taken for constructing data models was to gather various use-cases for data discovery, data retrieval, and data analysis, by interviewing astronomers for examples of usage and relying on the existing know-how of large data centers. From the collected ideas, we have tried to propose reconciling schemes for astronomical metadata description.

## 3. The data model landscape of the IVOA

Data modeling has been a central activity for the VO development as shown in the IVOA architecture document (Arviset and Gaudet, 2012). The interactions of the data modeling task principally lie in the definition of search parameters and representation of returned results in access protocols, so namely with the Data Access Layer Working Group (DAL WG). The class definitions elaborated by the DM WG can also feed the design of VO-aware applications by the Applications WG. The choice of a serialization format, to transport the modeled metadata, also involves the VOTable and the Semantics Working Groups. The ***ivoa.net*** (IVOA, 2014) document repository retains copies of the various data models available today. These have been designed for particular kinds of data products, first for space–time coordinates, then spectral datasets, 2D sky images, spectral lines, and data cubes. Most of the metadata used in astronomical protocols and applications are now derived from a stable set of data models as show in Table 1. These models are implemented and used by access protocols and client applications to effectively transport, visualize, transform, and interpret science observations.

## 4. Data modeling process

### 4.1. Metadata all around

Data modeling is focused on the metadata that describe the measurement values within an observation file: instrument name, file identifier, data provider, date of observation, date of publication, rights, position on the sky, field of view, instrument configuration, measurement quality, etc.

Historically this information was generally expressed in the FITS header keywords, with only a small set of standardized keyword labels for numerical data formats and WCS (World Coordinate Systems) location information. Most of the keywords used to express observation conditions, processing configuration, etc. do not obey a standardized vocabulary across the astronomical data centers.

Therefore a common, homogeneous, and structured representation of all these metadata was highly desirable in order to facilitate interoperability. Diverse use cases from protocol design in DAL WG, and from dataset handling in applications (Applications WG) helped to clarify usage of metadata and to sort out categories and roles for different pieces of metadata.

The concept of object oriented description appeared to be an adequate mechanism to represent various categories of metadata and group them logically. For instance, classes for Dataset, Curation, and Identification had been designed early in the Resource Metadata standard and organized in a tree-like structure as an XML schema[3] in VOResource and VODataService (Plante et al., 2010). These concepts, first stated in the Registry WG, are valid throughout the VO and are reused as key building blocks in other data models.

Coordinates (positional, temporal, spectral) are central in astronomy and have been modeled in detail, together with the various Coordinate Systems in the Space–Time Coordinates (STC)specification (Rots, 2011).

UML has been used since 2003 to express relationships between different concepts using class diagrams, and specifically to define each class and its attributes in detail. A text description is necessary to explain the properties of each class or attribute. It is provided along with each UML class diagram in the IVOA standard documents.

### 4.2. First steps to resolve metadata heterogeneity

Up to the 2000s, most astronomical data providers and reference archives used to store and distribute their observational datasets and the metadata attached to it following their own logic and policy. Each archive had its own interface. In order to homogenize the descriptions of source catalogs, the VOTable format has been proposed right at the beginning of the VO experience, to encode data and metadata in a tabular format. This very common data structure allows storage of all kinds of lists or collections with rows to store individuals, like sources, datasets, events, etc. and with columns representing properties measured or assessed for such individuals. The VOTable specification (Ochsenbein et al., 2011) defines basic XML elements such as FIELD, PARAM, and GROUP with attributes that qualify them. Among these qualifiers, two of them have led to the first steps of vocabulary standardization in the VO project: 'ucd' for the semantic content and 'unit' for expressing the units used for a column value.

The Unified Column Descriptor (UCD) (Preite Martinez et al., 2011) specifies a controlled vocabulary for the classification of the physical quantities exposed as a value in a column of a table. The collection of UCD terms covers most of the kind of measurements recorded in catalogs and observations in general. In the Data modeling effort, UCD words are used to add semantic value to some classes' attributes, like for instance to disentangle various kinds of flux measurements.

Another specification on which VO data modeling is based on, is the syntax definition of units strings for all measurement or metadata exposed in the VO system. Attributes describing such units in a VO Data Model should conform to the VO Units specification (Derriere et al., 2014) which gives the rules to compose a unit expression.

### 4.3. Metadata's scope and data model coverage

The DM WG has produced IVOA data models that are as comprehensive as possible with respect to their use cases, and

[3] See schemata at http://www.ivoa.net/xml/index.html.