Data modeling in the Virtual Observatory Framework

Mireille Louys
CDS, Strasbourg Observatory
ICube Laboratory, Strasbourg University, FR
A collaborative effort thanks to

Jesus Salgado, Carlos Rodrigo, Pedro Osuna, Mark Allen, Mireille Louys, Jonathan McDowell, Deborah Baines, Jesus Maiz Apellaniz, Evanthia Hatziminaoglou, Sebastien Derriere, Gerard Lemson, Laurent Bourges, Miguel Cervino, Claudio Gheller, Norman Gray, Franck LePetit, Benjamin Ooghe, Rick Wagner, Herve Wozniak, Igor Chilingarian, Marie-Lise Dubernet, Evelyne Roueff, Matteo Guainazzi, Enrique Solano, Joe Mazzarella, Raffaele D’Abrusco, Tamas Budavari, Markus Dolensky, Inga Kamp, Kelly McCusker, Doug Tody, Pavlos Protopapas, Arnold Rots, Randy Thompson, Frank Valdes, Petr Skoda, Bruno Rino, Jim Cant, Omar Laurino, Patrick Dowler, Daniel Durand, Laurent Michel, François Bonnarel, Mark Cresitello-Dittmar, Mark Markus Demleitner, Tom Donaldson, Patrick Dowler, Matthew Graham, Jose Enrique Ruiz, Kristin Riebe, Michèle Sanguillon, Mathieu Servillat, Catherine Boisson
IVOA Working Groups / Interactions

- Data Model
  - VOTable
    - Transport: XML
  - VOEvent IG
  - Theory IG
  - Registry
    - Localise resources and services
  - Grid/WebServices
    - WS FrameWork
    - Grid Computing facilities
  - Applications
    - Visualise, compare, compute
  - DataAccess Layer (Protocols)
    - Exchange, circulate
  - VOQL
    - Query Language for astronomical databases
  - Semantics
    - Unified Content Descriptor, Vocabulary, Units

- Data Exchange, circulate
Metadata modeling goals in the VO context

Describe metadata for all datasets exchanged in the astronomical community in an **homogeneous** way.

**Sustains the interoperability** objective of the Virtual Observatory

Based on the requirements:

- **For users**: 
  To be able to ask the same question/query to various astronomical data bases, select the results and then retrieve data.
  Search seamlessly for observational data at all wavelengths and for all instruments.

- **For data providers**: 
  Expose and distribute more data with a **standardized metadata description**. 
  In core in the DB structure or as an interoperability layer on existing architecture (DB view).
What is metadata in our context?

- Organize
- Describe at the appropriate level of details
- Each use-case sets its own quality requirements and criteria
- Describe all properties on each physical axis of the data: spatial, temporal, spectral, polarization, etc.
Following the principle of Object Oriented Design:

- Describe the **responsibilities** and **properties** of the metadata involved for data produced by an observation or a simulation process.
- A dialog between data providers and users who came up with:
  - **Root concepts** qualifying astronomical data
  - A **vocabulary** (list of terms) based on existing practises (FITS, bibliographic services, archives contents, interviews from astronomers and data providers ..)
  - **Logical structure** showing the dependencies and relations between all pieces of metadata
  - A general schema for metadata representation built from commonalities and specificities used at various data providers archives
## General building blocks models

<table>
<thead>
<tr>
<th>Metadata features</th>
<th>Data model name &amp; version</th>
<th>Year</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astronomical Space Time Coordinates</td>
<td>STC v 1.33</td>
<td>2007</td>
<td>REC</td>
</tr>
<tr>
<td></td>
<td>STC v 2.0</td>
<td>2016</td>
<td>WD</td>
</tr>
<tr>
<td>Physical axis description and properties</td>
<td>Characterization v1.13</td>
<td>2008</td>
<td>REC</td>
</tr>
</tbody>
</table>

REC: IVOA Recommendation  
PR: Proposed Recommendation  
WD: Working Draft
<table>
<thead>
<tr>
<th>Metadata features</th>
<th>Data model name &amp; version</th>
<th>Year</th>
<th>Status</th>
<th>Protocol</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectral Line Transitions</td>
<td>Simple Spectral Line</td>
<td>2010 REC</td>
<td>SLAP</td>
<td>VOSpec, SPLAT-VO</td>
<td></td>
</tr>
<tr>
<td>1D Spectrum, Light Curves</td>
<td>Spectrum v1.0</td>
<td>2007 REC</td>
<td>SSA 1.0</td>
<td>SPLAT-VO, IRIS</td>
<td></td>
</tr>
<tr>
<td>1D Spectrum, Light Curves</td>
<td>Spectrum v1.1</td>
<td>2011 REC</td>
<td>SSA 1.1</td>
<td>SPLAT-VO, IRIS</td>
<td></td>
</tr>
<tr>
<td>SED, Photometric Points, Time series, Multi-segment 1D spectrum</td>
<td>Spectral v2.0</td>
<td>2014 PR</td>
<td>SSA 1.1</td>
<td>SPLAT</td>
<td></td>
</tr>
<tr>
<td>Observational dataset Core Components</td>
<td>ObsCore v1.0</td>
<td>2011 REC</td>
<td>TAP 1.0, SIAv2</td>
<td>TAPHandle, TOPCAT</td>
<td></td>
</tr>
<tr>
<td>(All data products for global discovery)</td>
<td>ObsCore v1.1</td>
<td>2016 PR</td>
<td>TAP 1.0</td>
<td>TAPHandle, TOPCAT</td>
<td></td>
</tr>
<tr>
<td>Photometric calibration</td>
<td>Photometry v1.0</td>
<td>2013 REC</td>
<td>SPLAT-VO, CDS Sed browser, SVO Filter Profile Service</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All observation datasets (Fine grain description)</td>
<td>DataSet Metadata v1.0</td>
<td>2015 WD</td>
<td>SPLAT-VO/time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N-D cubes; pixelated images, sparse data</td>
<td>NDCube v1.0</td>
<td>2015 WD</td>
<td>SPLAT-VO/time</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Other Metadata

<table>
<thead>
<tr>
<th>Metadata features</th>
<th>Data model name &amp; version</th>
<th>Year</th>
<th>Status</th>
<th>Protocol</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOEvent: transients observations</td>
<td>VOEvent v2.0</td>
<td></td>
<td>REC</td>
<td>VTP 1.0</td>
<td>Service embedded desc .</td>
</tr>
<tr>
<td>Simple time series</td>
<td>Simple time series v1.0</td>
<td>2014</td>
<td>Note</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Simulation Data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simulations, data and code description</td>
<td>Simulation v1.0</td>
<td>2012</td>
<td>REC</td>
<td>SimDAL</td>
<td>DEUVO, Meudon PDR code, MilleniumDB, etc.</td>
</tr>
<tr>
<td>Micro simulations, Implementations of SimDM</td>
<td>Implementations of the Simulation DM v1.0</td>
<td>2012</td>
<td>Note</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Provenance metadata</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Datasets generation process, Progenitors</td>
<td>Provenance DM v1.0</td>
<td>2016</td>
<td>WD</td>
<td>TBD</td>
<td></td>
</tr>
</tbody>
</table>
• The DM fields arrangement contains the dependencies between different pieces of metadata

• Can be expressed as
  – UML class diagram
  – Mind maps
  – Tables
  – Lists

Suitable the most for interpretation by humans (graphs) or by machine (tables, lists, XML documents)
<table>
<thead>
<tr>
<th>column</th>
<th>unit</th>
<th>ucd</th>
<th>uctype</th>
<th>dataType</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>obs_publisher_did</td>
<td></td>
<td>meta.ref;url;meta.curation</td>
<td>obscore:Curation.PublisherDID</td>
<td>VARCHAR</td>
<td>publisher dataset identifier</td>
</tr>
<tr>
<td>obs_collection</td>
<td></td>
<td>meta.id</td>
<td>obscore:DataID.Collection</td>
<td>VARCHAR</td>
<td>short name for the data collection</td>
</tr>
<tr>
<td>facility_name</td>
<td></td>
<td>meta.id;instr.tel</td>
<td>obscore:Provenance.ObsConfig.Facility.name</td>
<td>VARCHAR</td>
<td>telescope name</td>
</tr>
<tr>
<td>instrument_name</td>
<td></td>
<td>meta.id;instr</td>
<td>obscore:Provenance.ObsConfig.Instrument.name</td>
<td>VARCHAR</td>
<td>instrument name</td>
</tr>
<tr>
<td>obs_id</td>
<td></td>
<td>meta.id</td>
<td>obscore:DataID.observationID</td>
<td>VARCHAR</td>
<td>internal dataset identifier</td>
</tr>
<tr>
<td>dataproduct_type</td>
<td></td>
<td>meta.id</td>
<td>obscore:ObsDataset.dataProductType</td>
<td>VARCHAR</td>
<td>type of product</td>
</tr>
<tr>
<td>calib_level</td>
<td></td>
<td>meta.code;obs.calib</td>
<td>obscore:ObsDataset.calibLevel</td>
<td>INTEGER</td>
<td>calibration level (0,1,2,3)</td>
</tr>
<tr>
<td>obs_release_date</td>
<td></td>
<td>time.release</td>
<td>obscore:Curation.releaseDate</td>
<td>TIMESTAMP</td>
<td>timestamp of date the data becomes publicly available</td>
</tr>
<tr>
<td>target_name</td>
<td></td>
<td>meta.id;src</td>
<td>obscore:Target.Name</td>
<td>VARCHAR</td>
<td>name of intended target</td>
</tr>
<tr>
<td>s_region</td>
<td>deg</td>
<td>phys.outline;obs.field</td>
<td>obscore:Char.SpatialAxis.Coverage.Support.Area</td>
<td>REGION</td>
<td>region bounded by observation</td>
</tr>
<tr>
<td>s_resolution</td>
<td>arcsec</td>
<td>pos.angResolution</td>
<td>obscore:Char.SpatialAxis.Resolution.refval.value</td>
<td>DOUBLE</td>
<td>typical spatial resolution</td>
</tr>
<tr>
<td>s_xel1</td>
<td></td>
<td>meta.number</td>
<td>obscore:Char.SpatialAxis.numBins1</td>
<td>BIGINT</td>
<td>dimensions (number of pixels) along one spatial axis</td>
</tr>
<tr>
<td>s_xel2</td>
<td></td>
<td>meta.number</td>
<td>obscore:Char.SpatialAxis.numBins2</td>
<td>BIGINT</td>
<td>dimensions (number of pixels) along the other spatial axis</td>
</tr>
</tbody>
</table>
- Derived from UML modeling
- Expressed following the VO-DML meta-model
- Describes all datamodel classes, attributes and relations in a dedicated XML document
- On going specification in proposed recommendation phase
Interaction aspect / server side
Interaction aspect /client side

Data Model
- IVOA standard document
- Schema
  - XML schema
  - VO-DML desc.
- Mapping rules

Documentation
- Annotation

Object Design
- Meaning types, units, rules, etc.

Client Application
- VO Query

Data provider Server
- VO Response
- VOTable document
  - XML, JSON
Dataset Metadata (abstract DM)

Data Dataset DataID Target Curation Characterisation

Spectral DM

ObsCore DM

ND-Cube DM

SED
Spectrum
TimeSeries
ObsDataset
ND-Cube

transported by

Dataproduct Class

defines

Serialised Instance

projection

Protocol

Data models in action

Dataproduct Class

uses concepts from

extends

protocol

Datamodel

Dataproduct Class

Data models in action

Dataproduct Class

Datamodel

Dataproduct Class

Datamodel

Dataproduct Class

Datamodel

Dataproduct Class

Datamodel

Dataproduct Class

Datamodel

Dataproduct Class

Datamodel

Dataproduct Class

Datamodel

Dataproduct Class

Datamodel

Dataproduct Class

Datamodel

Dataproduct Class

Datamodel

Dataproduct Class

Datamodel
Lessons learnt (development strategy)

- Incremental development
- Gathering use-case has proven to be effective
- Discussions with data providers can be cumbersome but are essential
- Testing in real context with applications, protocols and data model update sets up a positive development feedback
- It was challenging to adjust the granularity level:
  - Details specific to particular data archives are not covered, but the common interoperable description layer works
Lessons learnt (technical aspects)

- Easy to prototype, but long and painful to get everything consistent and persistent in a large reference picture.
- Important to find the appropriate scope for data modeling:
  - Common features → mandatory data model fields
  - Specific features → optional features
  - We target interoperability and not exhaustivity
- Data model compliance checks improved when we can validate serialisations (owing to the Operations WG)
• VO data models offer a rich set of data models describing most of observation and simulation data products
• A good learning curve for VO developers thanks to the motivation of the actors
• Data models build up the underlying semantic and structural layer that binds together the VO framework
• Join us at IVOA meetings or contact us on dm@ivoa.net
Thanks for your attention Thanks to all people involved
A collaborative effort thanks to

Jesus Salgado, Carlos Rodrigo, Pedro Osuna, Mark Allen, Mireille Louys, Jonathan McDowell, Deborah Baines, Jesus Maiz Apellaniz, Evanthia Hatziminaoglou, Sebastien Derriere, Gerard Lemson, Laurent Bourges, Miguel Cervino, Claudio Gheller, Norman Gray, Franck LePetit, Benjamin Ooghe, Rick Wagner, Herve Wozniak, Igor Chilingarian, Marie-Lise Dubernet, Evelyne Roueff, Matteo Guainazzi, Enrique Solano, Joe Mazzarella, Raffaele D’Abrusco, Tamas Budavari, Markus Dolensky, Inga Kamp, Kelly McCusker, Doug Tody, Pavlos Protopapas, Arnold Rots, Randy Thompson, Frank Valdes, Petr Skoda, Bruno Rino, Jim Cant, Omar Laurino, Patrick Dowler, Daniel Durand, Laurent Michel, François Bonnarel, Mark Cresitello-Dittmar, Mark Markus Demleitner, Tom Donaldson, Patrick Dowler, Matthew Graham, Jose Enrique Ruiz, Kristin Riebe, Michèle Sanguillon, Mathieu Servillat, Catherine Boisson