EuroVO-AIDA

Euro-VO Astronomical Infrastructure for Data Access

D7.1

–

Intermediate report on Data Access Protocols and Data Models and associated prototypes

Final version

Grant agreement no: 212104

Combination of Collaborative Projects & Coordination and Support Actions

DOCUMENT INFORMATION

Project

Project acronym:	EuroVO-AIDA
Project full title:	Euro-VO Astronomical Infrastructure for Data Access
Grant agreement no.:	212104
Funding scheme:	Combination of Collaborative Projects & Coordination and Support Actions
Project start date:	01/02/2008
Project duration:	30 months
Call topic:	INFRA-2007-1.2.1 Scientific Digital Repositories
Project web sites:	http://www.euro-vo.org/pub/general/intro.html
	http://cds.u-strasbg.fr/twikiAIDA/bin/view/EuroVOAIDA/WebHome

Document

Deliverable number:	D7.1
Deliverable title:	Intermediate report on Data Access Protocols and Data Models and associated prototypes
Due date of deliverable:	April 2009
Actual submission date:	04 September 2009
Authors:	ESA
Work Package no.:	WP7-JRA2
Work Package title:	Data Access Protocols and Data Models
Work Package leader:	ESA
Lead beneficiary:	ESA
Dissemination level:	PU
Nature:	Report + Other
No of pages (incl. cover):	19

TABLE OF CONTENT

1. OBJECTIVES OF WP7 GROUP 4

2. IMPLEMENTATION STRATEGY FOR WP7 GROUP 4

2.1. WP7-T1 Asynch DAL 5

2.2. WP7-T2 SED Library 5

2.3. WP7-T3 Footprint DAL 5

2.4. WP7-T4 Generic Data Set 5

2.5. WP7-T5 Associations in S*AP response 5

2.6. WP7-T6 Photometry Data Model 5

2.7. WP7-T7 Radio Cubes 5

2.8. WP7-T8 ADQL Library 5

2.9. WP7-T9 Source Catalogue Data Model 5

2.10. WP7-T10 Observation Provenance and Characterisation Data Models 6

2.11. WP7-T11 Units 6

3. DETAILED WORK PERFORMED 6

3.1. WP7-T1 Asynch DAL 6

3.2. WP7-T2 SED library 6

3.3. WP7-T3 Footprint effort and integration in DAL 7

3.4. WP7-T4 Generic Data Set discovery services definition 8

3.5. WP7-T5 Associations and Extensions in S*AP 9

3.6. WP7-T6: Photometry Data model and services 10

3.7. WP7-T7 Radio cubes 13

3.8. WP7-T8 ADQL Library 14

3.9. WP7-T9: Source Catalogue Data Model 14

3.10. WP7-T10 Observation – Provenance – Characterisation DM 15

3.11. WP7-T11 Units 17

ACRONYM LIST 16

1. OBJECTIVES OF WP7 GROUP

The idea of Work Package 7 – Joint Research Activity 2 is to perform R&D studies for Data Access, Data Model and VObs Query Language protocols and standards. WP7 is led by ESA with contributions from the other WP7 participants, in particular of CNRS and UK for the data modelling and some data access layer aspects, of ESO and of CNRS, INAF, INTA and UHEI for theory standards.

In particular that includes:

- Studying the DAL, DM, VOQL standards. Assessing possible technologies to implement these standards
- Participation to the IVOA WGs discussions, mail forum, teleconferences
- Participation to writing the standards specifications document of these DAL, DM, VOQL
- Participation to the IVOA Technical forum to present the findings and implementations of these DAL, DM, VOQL standards
- Writing prototype implementation these DAL, DM, VOQL standards onto real data holdings from European Data Centres
- If applicable, publish these implementations into the Euro-VO Registry of VObs resources
- Coordinate with EuroVO-AIDA Service Activities for these (prototype) implementations of these DAL, DM, and VOQL standards to be consumed by VObs tools (in liaison with WP3 and WP4)

Note that other standards and protocols are addressed by WP6 – JRA1 "*Evolution of VObs interoperability standards*", with cross-links with WP7. The bulk technical work on IVOA standards is undertaken in JRA1 and JRA2, but the co-operation between partners is a Networking Activity undertaken in WP2-3 (EuroVO-AIDA Technology Forums) and in WP2-4 (networking with the other members of the IVOA). This report formally covers the first Cycle of EuroVO-AIDA project activities, until April 2009. The activities have been discussed at the Second and Third Euro-VO Technology Forums, which were held in Cambridge, UK, on 29th September - 2nd October 2008, and in Strasbourg, France, on 16th - 18th March 2009. Cycle 1 activities were assessed during the Mid-Term EuroVO-AIDA Board meeting, held in Paris on 29th June 2009, taking into account discussions at the Strasbourg IVOA Interoperability meeting (24th - 29th May 2009).

All IVOA documents and standards can be found at http://www.ivoa.net/Documents/.

2. IMPLEMENTATION STRATEGY FOR WP7 GROUP

In order to effectively implement the aforementioned objectives, the work was subdivided into different so called "Tasks", each of those tasks being led by one of the WP7 working partners.

The following tasks were identified.

2.1. WP7-T1 Asynch DAL

The objective of this task is to study the Asynchronous Query mechanisms for the future protocols (TAP, SIAv2, etc.).

2.2. WP7-T2 SED Library

The purpose of this WP7 task is to facilitate VO adoption for Spectra. A small library, to be integrated in the server side by a data centre, should translate "native" fits to IVOA Spectrum Data Model fits format. Hence any clients able to understand this standard fits format can handle spectra stored in a "native" fits format.

2.3. WP7-T3 Footprint DAL

The purpose of this task is to study how to integrate properly footprint information within the available European resources and to bring to the IVOA ideas on how to homogenize its access.

2.4. WP7-T4 Generic Data Set

The purpose of this task is to study how to represent and access Data Sets of generic type, i.e., not restricted to images or spectra. This can be achieved in several ways, and it is the objective of this task to investigate what the best approach is.

2.5. WP7-T5 Associations in S*AP response

The purpose of this task is to investigate how so called associations (i.e., multiple images, Echelle spectra, etc.) given in response to S*AP protocol calls can be handled by defining a mechanism to allow certain hierarchical structure representation within the S*AP response.

2.6. WP7-T6 Photometry Data Model

The purpose of this task is to try and define a proper model for Photometry -which is currently under defined within the IVOA- and then bring those ideas to the IVOA for eventual endorsement.

2.7. WP7-T7 Radio Cubes

The purpose of this task is to better define radio cubes and their access within the IVOA.

2.8. WP7-T8 ADQL Library

The purpose of this task is to study the conversion between the ADQL language and other available standards, predominantly SQL. Conversion libraries should ideally be created to seamlessly translate from one language to the other.

2.9. WP7-T9 Source Catalogue Data Model

The purpose of this task is to define a Source Catalogue Data Model that would help in the discovery of Source Catalogues with the help of the near future TAP protocol.

2.10. WP7-T10 Observation Provenance and Characterisation Data Models

The purpose of this task is to continue the work being done on characterizing observational data, both from the point of view of the data and from the point of view of its provenance.

2.11. WP7-T11 Units

The purpose of this task is to define a standard for the usage of units within the IVOA globally that could be proposed eventually for endorsement.

3. DETAILED WORK PERFORMED

One important result is that the Astronomical Data Query Language (ADQL) has become an IVOA Recommendation on 30th October 2008 under ESA leadership. The latest version of the document was presented at the Trieste Interoperability meeting in May 2008, where final minor points were addressed. The mandatory period of Request For Comments was initiated some time afterwards, when the pertinent changes had been introduced. The document was set for approval before the Baltimore Interoperability meeting, where some final touches had to be done. The Executive committee finally approved the document on 30th October 2008. The final document can be found at: http://www.ivoa.net/Documents/cover/ADQL-20081030.html.

3.1. WP7-T1 Asynch DAL

No work has been done during this period on this task.

3.2. WP7-T2 SED library

The IVOA Simple Spectra Access Protocol (SSAP) and its underlying data model (SpectrumDM) fully support the kind of spectra (FITS) data formats that are normally used within the Space Astronomy community (binary FITS tables). Instead, the Ground-based community has come up in the years with many different and incompatible ways of storing spectra in a FITS file. As stated in the SSA standard:

> [...]*Spectrum datasets may conform to a standard data model defined by SSA, or may be native spectra with custom project-defined content.* [...]

Spectral data is generally stored externally to the VO in a format specific to each spectral data collection; currently there is no standard way to represent astronomical spectra, and virtually every project does it differently. Hence spectra may be actively mediated to the standard SSA-defined data model at access time by the service, so that client analysis programs do not have to be familiar with the idiosyncratic details of each data collection to be accessed.

Though, many Ground-based observatories have adopted a so-called 1D-image format, less descriptive than the IVOA defined binary FITS table format. To bridge the gap ground-based data providers should convert their 1D-image to the IVOA suggested format.

This task proposes to study the possibility to create a generic conversion library that might help the data providers to actually transform their original ("native") ground based FITS format in a fully SpectrumDM compliant format.

Current Status

At ESO a first effort is based on the "natural" native FITS format most of the ground based observatories are using world-wide. A library is being developed to convert the ground-based typical spectral format (1D image) into the SpectrumDM suggested FITS format. A CGI script to test the library is already available, and will be soon made available through the ESO SSA operational service. Later on, based on this concrete experience, the library will be made available to other Data Centres.

3.3. WP7-T3 Footprint effort and integration in DAL

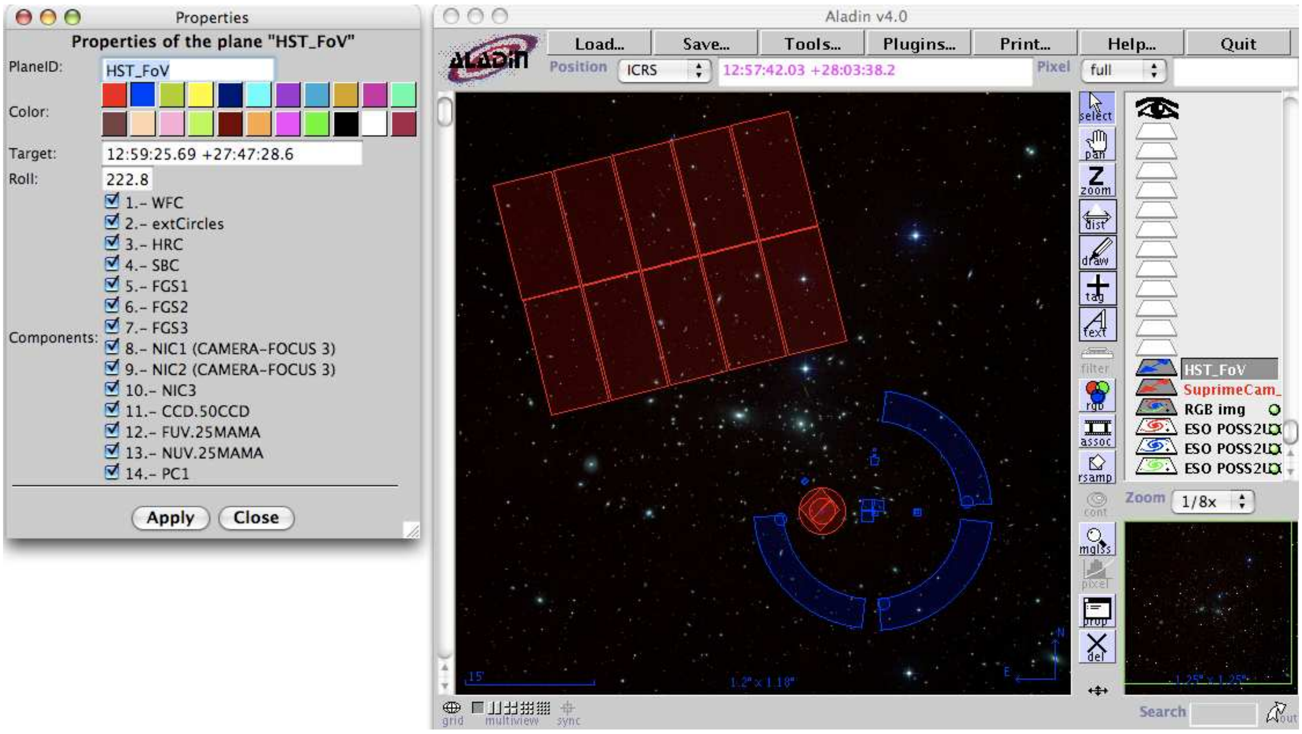
A online graphical editor has been developed, allowing one to build up Field Of View (FOV) in the format described by the IVOA note called *Footprint Overlay Specification* available on the IVOA site. This tool was presented in the Baltimore IVOA Interoperability meeting in October 2008 ([http://www.ivoa .net/internal/IVOA/InterOpOct2008DAL/CDS-footprint-editor.pdf](http://www.ivoa.net/internal/IVOA/InterOpOct2008DAL/CDS-footprint-editor.pdf)).

The Hubble Legacy Archive service provides now footprint of Hubble Images in this format. This is a collaboration with T. Donaldson from the Space Telescope Science Institute (available in next Aladin version).

In the meantime, we had further discussions with the members of the National Virtual Observatory (NVO – the US-VO project) involved in the FootPrint service IVOA proposal (T. Budavari, G. Greene). Tests examples have been provided for the description of an observation sketching out some parts of the Observation Data Model. They reuse the IVOA Characterization Data Model and include footprint information as the spatial support element using an Space-Time Coordinates (STC) AstroCoordArea class.

The Footprint service interacts at the STC structure level which is common to both approaches. This Observation Container prototype, including dataset ID and footprint information, was successfully tested for interoperability with the NVO footprint service. All this was reported at the IVOA Baltimore Interoperability meeting. A few revisions of the XML document examples have slightly improved the prototype in the meantime. Collaborative work is going on.

Example of instrument footprints displayed in Aladin (blue: Hubble Space Telescope; red: SuprimeCam on the Japanese Subaru telescope



3.4. WP7-T4 Generic Data Set discovery services definition

This package is a generic DAL protocol for all kind of observation data.

It is clear that generic dataset interface has to be defined in relationship with TAP and SIA/SSA protocols.

TAP effort has now made large progress and will go to PR rather soon. SIA2 is now a top priority and work on Generic dataset should not be separated from this SIA2 effort.

Discussions on the SIA2 interface include "QUERY" parameters, access to cubes, with cutouts and transformations (including resampling), output parameters with their utypes, etc, inclusion of DAL associations and extensions.

The Generic dataset access protocol will include a common profile to encapsulate the existing IVOA simple access protocols and provide advanced capabilities such as ADQL queries, descriptions of metadata according to the full Observation data model, and access modes to heterogeneous complex data.

A version of the SIA2 preliminary draft interface is in preparation for end of June. A version of SIA2 working draft is planned to be circulated for 3rd August 2009. Then the generic Dataset protocol will be designed following the vision developed in *DAL architecture document* (http://www.ivoa.net/internal/IVOA/SiaInterface/DAL2_Architecture.pdf - hereafter DAL2 designates second generation Data Access Layer) and formalized in a new document.

3.5. WP7-T5 Associations and Extensions in S*AP

Clients and servers use associations and extensions in a marginal way in the S*AP context. Discussions were held between SIA2 core members at IVOA level to analyse the reasons and find possible solutions. Some agreements obtained at IVOA level within this task:

1 - Lack of examples in IVOA specifications:

Examples of possible associations/extensions will be provided. As all the cases cannot be foreseen, the examples will be published with a disclaimer. A possible alternative location for the examples could be the DAL2 architecture document.

2 - Description of association or extension in IVOA S*AP protocols:

A general mechanism for associations is already present in the current DAL2 (SSAP) specification.

This description includes:

- *Association.Type:* Type of association
- *Association.ID:* Unique ID identifying the association instance
- *Association.Key:* Unique key different for each element association

Only a predefined case (Multiformat association) is described, where several records refer to the same dataset which is available in several different output data formats.

A description of a general mechanism is also included in the SSAP but the level is too abstract to be used by a final software developer for the server or, in particular, for the client side. A fine grain description is still needed. There are not many details about extensions in the specification.

As for the previous point, some examples will be provided, with a disclaimer (not all the cases can be foreseen). Possible a better location for this discussion could be the DAL2 architecture document.

3 - Use of association or extension for different use cases:

For some use cases, both the extension and the association approaches are valid. A common rule is that heterogeneous data should be represented as an extension.

We identified a list of use cases to be analyzed:

- Complex data: multi-images, MOS, Echelle, more for Generic data Set?

- Complex metadata: Photometry, Provenance, characterization level 3 (FOVs) and 4 (variability maps)

- Echelle spectra:

  - It could be described either by association of multiple 1D spectra or use a multi-segment spectral representation. It this enough?
  - Usually the entire spectrum is also accessible in one of the links. The hierarchical structure is not defined in the association type (Association Type=hierarchical?) (Standard way to link the records including the hierarchical information?)

- Transmission curves/response matrixes linked in the SSAP response to SED: Records associated using association mechanism. Link to the curve/matrix using the extension mechanism. This is particularly important to include high energy spectra in the VO, so a review of the spectral data model is needed

- Images and sub-images: In SIAP, one image per observation + images for different energy bands or instrument. Example XMM images:

  - It could be done by association. Same problem as in the Echelle spectra to describe the hierarchical information. It can be also done as extension, although this could increase the level of complexity in the interpretation of the response for the client
  - Some multiband surveys examples. Check and select one as a reference implementation

- Multi-position query: Records associated using the association mechanism. Extension (if needed) to give target details

- Confidence Maps (McMahon): "a normalised inverse variance weight map defining the "confidence" associated with the intensity value in each pixel, this also encodes for hot pixels, bad pixels, dead pixels and so on". Provenance example

3.6. WP7-T6: Photometry Data model and services

Photometry data is one of the most important types of data to be included in the Astrophysical Virtual Observatory. However, the lack of services providing this type of data reflects that the photometry data has not been properly described in the VO context.

The present task has the following objectives:

- Propose a Data Model to be agreed in the EuroVO-AIDA context and propose it to the IVOA

- Understand and describe the use of synthetic photometry in the VO context

- Understand and describe how to publish photometry data in the VO context

Activities during current period:

- Meeting held between SVO/ESA-VO members

- Preliminary Photometry UML Data model produced. Task distribution working list created and model distributed to it

- Science use cases distributed to the list

- Plan of activities developed and distributed

- Distribution of preliminary Photometry UML for discussion (done)

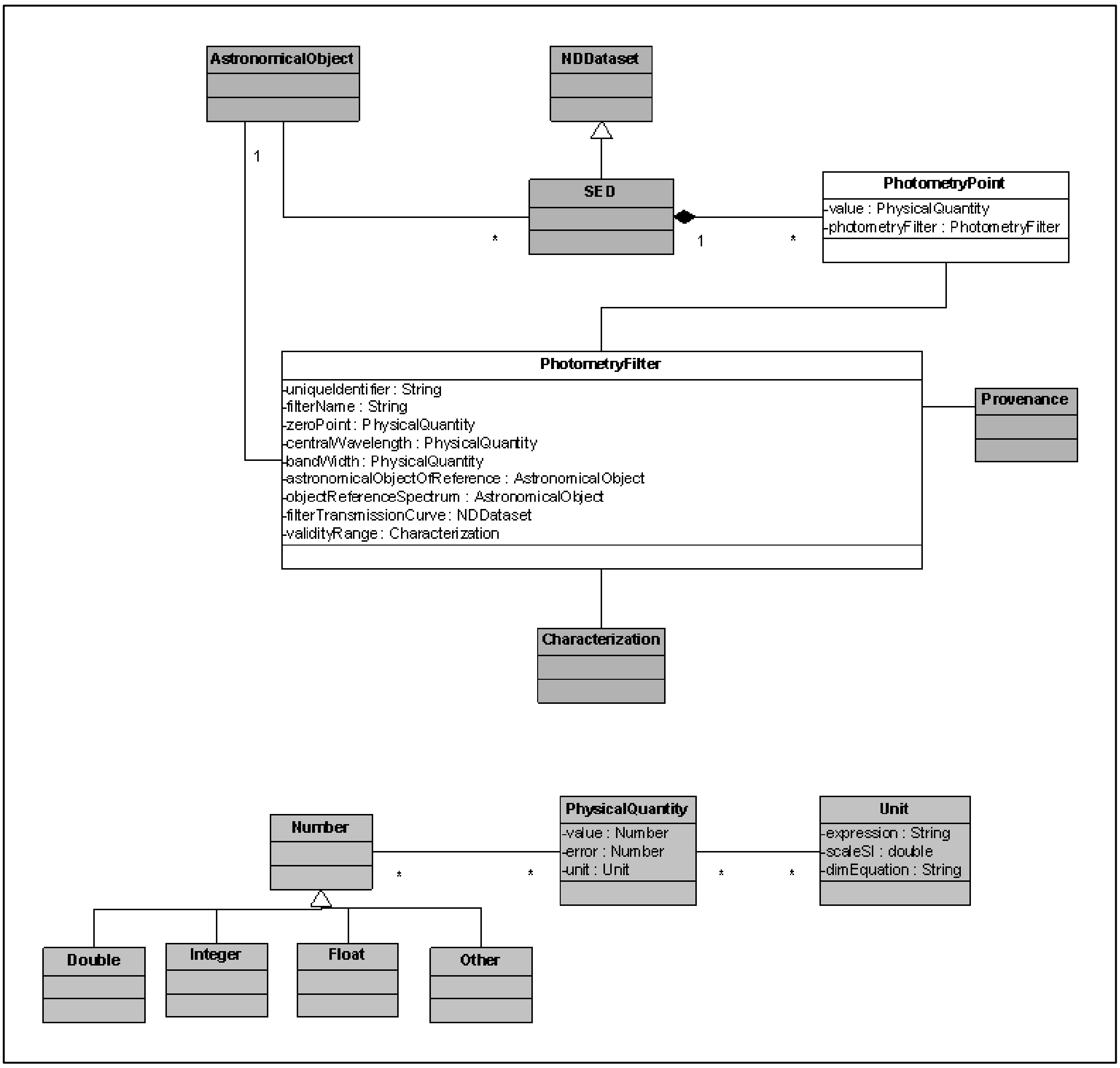
- Distribution of science use cases (done)

- Agreement on DM
- Creation of Profile Filter Service and interaction with VO applications (ongoing)
- Creation of Photometry data services and interaction with VO applications

➢ Model presented during the Euro-VO Technology Forum (March 2009)

➢ Discussions held during the meeting and changes implemented in the model

Photometry Data Model

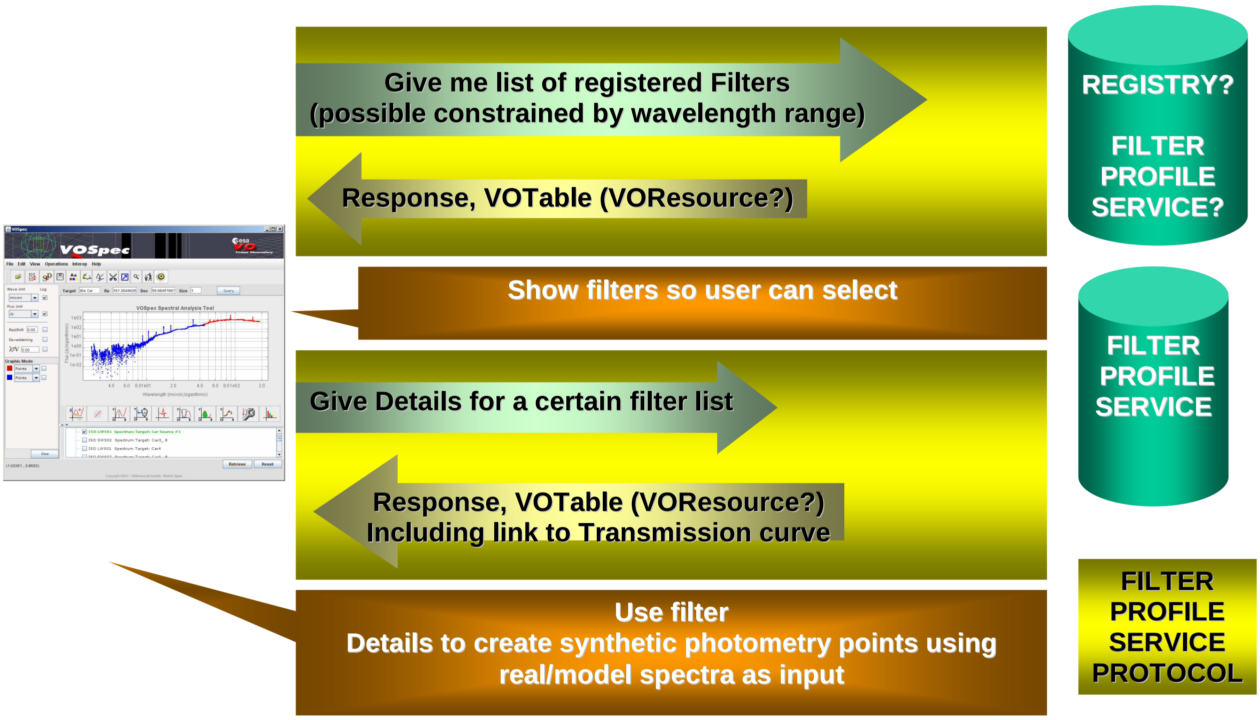


Proposal of protocol for synthetic photometry

In order to allow the creation of synthetic photometry from a real or theoretical SED, the following workflow has been suggested within the context of this task. A new protocol (Filter Profile Service Protocol), in line with the standard IVOA S*AP protocols, has been proposed.

The steps are as follows:

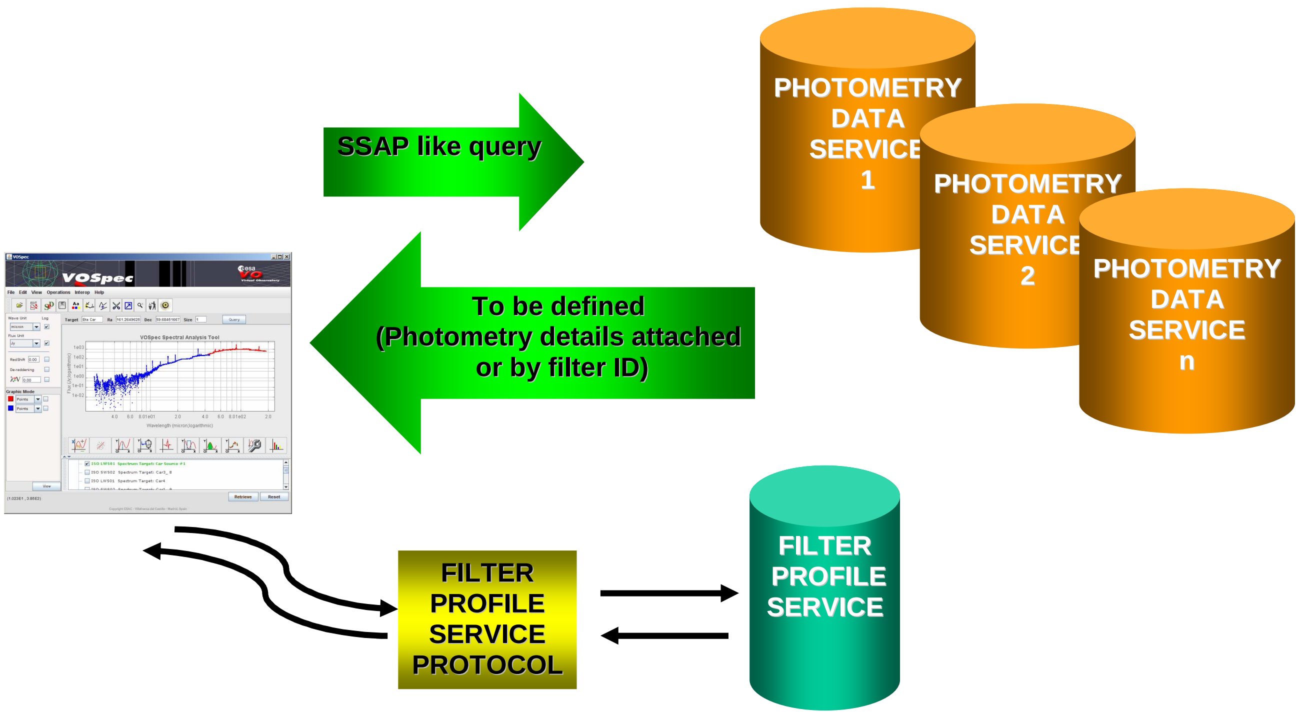
- Client application will ask Photometry Filter Profile Server for a list of registered photometry filters. This could be done at server level or a registry level (under discussion)
- Server will return the list of registered filters + unique identifier
- Client application will request for photometry filter details by using the unique identifier
- Photometry filter details will be returned in an agreed format (under discussion)
- Client application will use this photometry filter details to produce synthetic photometry



Proposal of interaction with SSAP

In order to allow the creation of SSAP server for photometry data, the following workflow has been suggested:

- Client do a standard SSAP request to the SSAP servers
- Servers will return a standard SSAP response. The data links would be in a to-be-agreed standard photometry data file format. Two approaches discussed:
    - Attached photometry filter information: For every photometry filter, the information of how to use the related data will be attached in the file
    - Detached photometry filter information: For every photometry filter, a unique identifier will be provided. If this is the case, an extra step will be needed to obtain the photometry filter information (see Filter Profile Server workflow described in previous point)
- Client application will display/use photometry data in combination with other VO published data



3.7. WP7-T7 Radio cubes

This task is looking at what is needed to support the publication and retrieval of multi-dimensional science-ready data which is in the form of a 'stack of planes'. This covers multi-frequency radio images and other spectral cubes in a similar format (but not raw IFU data, for example), and also data where the third axis is polarization parameter or time; there may even be more than three axes.

This will be encompassed within the SIAP v2 extension. The contributions from the EuroVO-AIDA WP members include:

- Producing a Note on polarization, also for presentation at the Interoperability meeting
- Summarizing briefly the requirements of data providers such as ALMA, e-MERLIN, HI single dish mosaics
- Investigating compatible formats in other wavebands, e.g. some ESO optical or IR data (SINFONI) may be in the form of suitable cubes
- Drawing up priorities for data retrieval standards
- Commenting on related aspects of SIAPv2

Note on polarization is available at:
http://www.ivoa.net/internal/IVOA/NotesOnPolarization/Note-Polarization-0.1-20090522.pdf.

3.8. WP7-T8 ADQL Library

No work has been done on this task during this period.

3.9. WP7-T9: Source Catalogue Data Model

The idea behind the Source Catalogue data model is to set up a framework where, by relying on the knowledge of a general catalogue data model, the astronomical queries could be built and passed to the services without having to know the peculiarities of the underlying models for each service. Therefore, the same query could be passed to different services, and each service would internally translate it to its own model transparently for the end user.

Therefore, making the *model* (Source Catalogue Data Model), the *query language* (ADQL) and the *access protocol* (TAP) work together, the Virtual Observatory would be enabled to send uniform queries across a distributed set of heterogeneous catalogue services; with the benefit that such framework could also be adopted by other models defined within the IVOA Data Model Working Group (http://www.ivoa.net/cgi-bin/twiki/bin/view/IVOA/IvoaDataModel).

Framework Architecture

This section details the proposed tiered architecture to fulfill the use case described previously: clients being able to send the same query to different and heterogeneous services by using the Source Catalogue Data Model.

Service

Different services must be available providing access to astronomical source catalogue data. Catalogue data is typically stored in relational database systems (RDBMS), where each service has its own data model implemented. Therefore, the services as such do not provide a uniform way to access them, requiring the protocol to define how this can be handled.

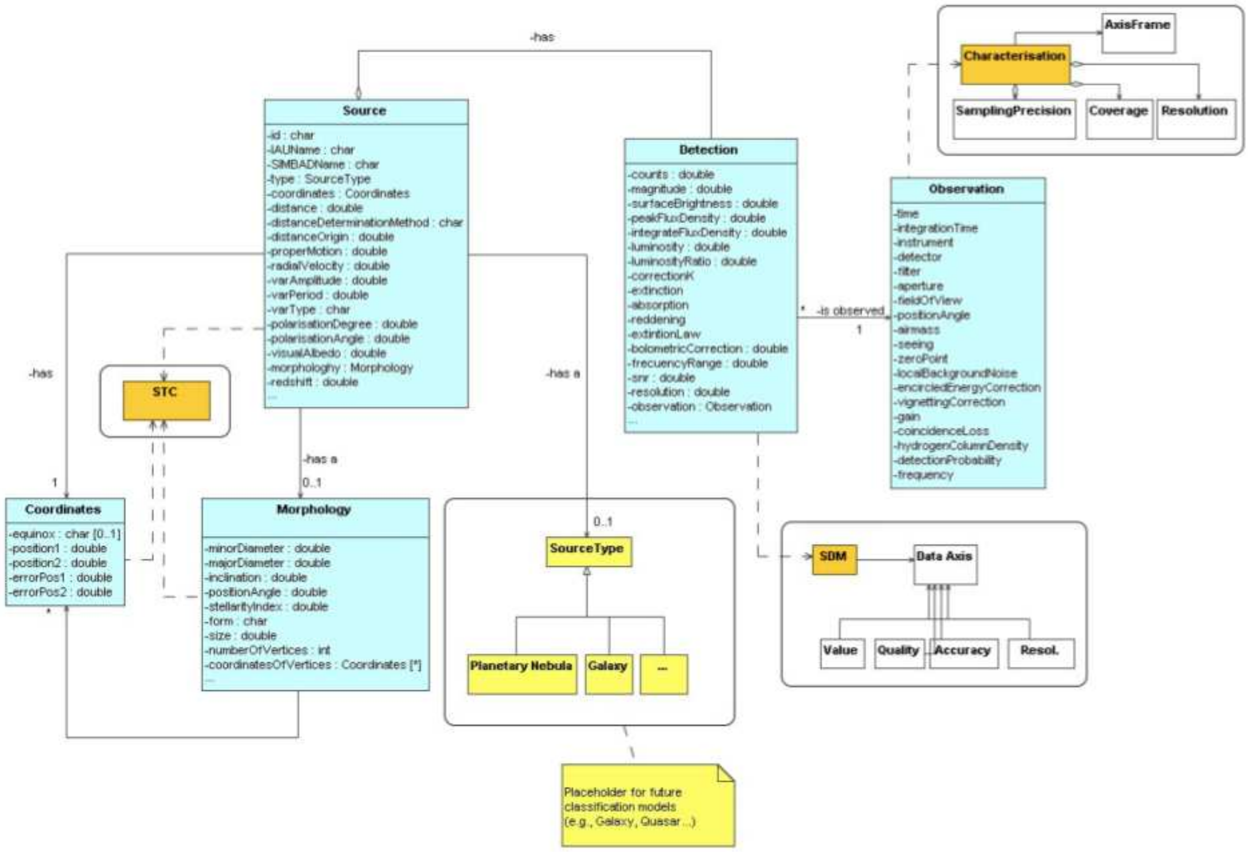
Protocol

On top of each service, there must be implemented a TAP layer defining how service features can be invoked to access tabular data including astronomical catalogues. This layer provides web access over standard HTTP protocol and accepts queries submitted in a specific format (See next section).

Query language

There must be defined explicitly, the payload that it is exchanged between the client and the service through the protocol. In this framework, the selected payload, compatible with the protocol, is the Astronomical Data Query Language.

The data model UML for the Source Catalogue at the time being is the following:



3.10. WP7-T10 Observation – Provenance – Characterisation DM

We are going to propose a slightly modified Characterisation data model version 2.0 and to propose a new list for Utypes, in order to:

- Adopt IVOA good practice for XML schema writing
- Take into account complex datasets
- Describe variability maps for coverage , resolution and sampling

Discussions about simplification of Data models have been going on.

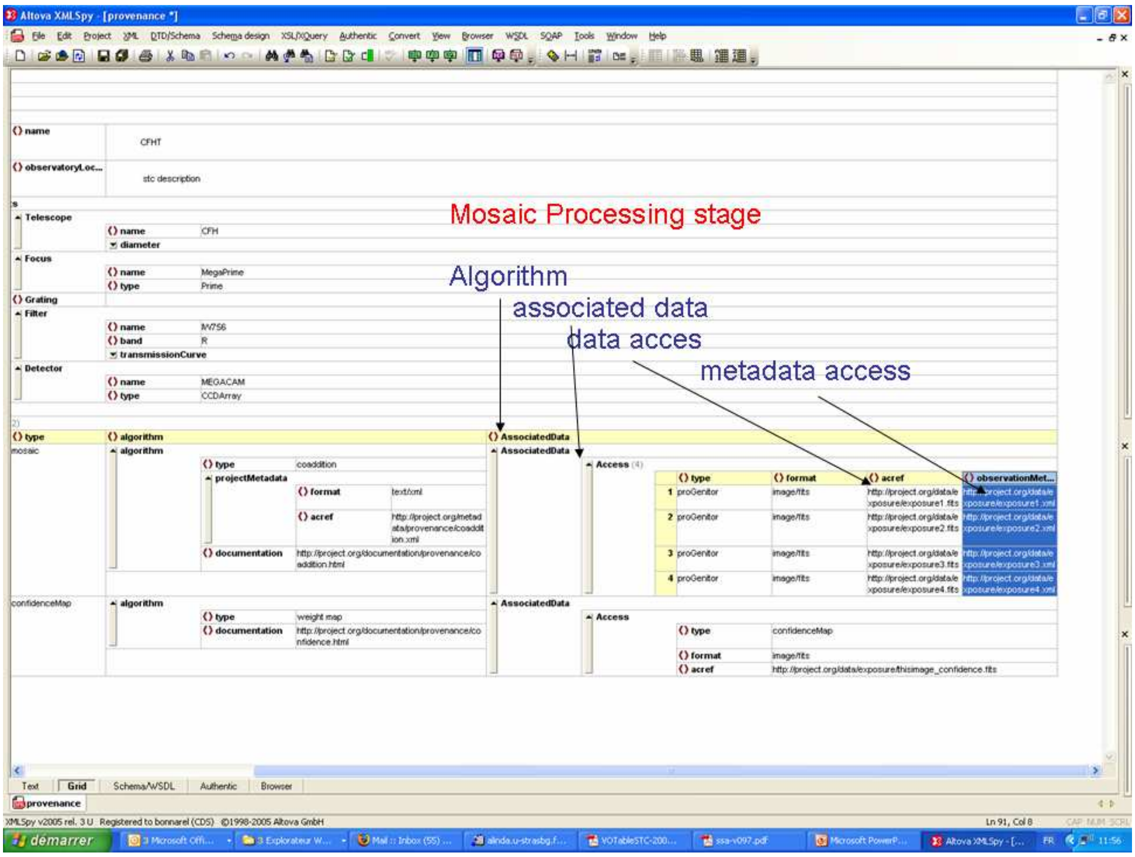
The Characterisation DM has been challenged in various advanced use-cases provided by scientists in order to describe complex data and experiment the fine grain detailed levels of the model, especially the Error and Variability maps.

Recently an example of error maps and resolution map for Sloan spectra has been built. In that case, solutions to provide pointers inside a multiextension dataset containing science data as well as ancillary data are proposed. Work is going on with I. Chilingarian.

Various XML instance documents have been sketched out for specific use-cases:

- Polarisation data
- Multiple CCD observations

<u>Example of Provenance for image mosaic (the model includes refernce to the algorithm used to build the mosaic)</u>



Observation Data Model

The Observation container has been designed including Provenance and Characterization Packages.

We developed various XML serialisation examples of Provenance: for a stacked image (association), for an observation with instrumental Provenance metadata: Telescope and Filter description.

An XML schema will be proposed in July 2009. A first draft will be presented at the Interoperability meeting in November 2009.

3.11. WP7-T11 Units

The need for a Units model and a Units controlled vocabulary has arisen from several reasons exemplified below:

1 - To understand m, meter, metre ... as the same thing

2 - To differentiate between mm (milli-metre) and m.m (m2)

3 - To distinguish between the use of m as a wavelength (= c/frequency) and m as a distance e.g. 1 AU = 1.499 1011 m

4 - To create combinations of units intelligently e.g. J.s-1 = W

5 - To translate between SI prefixes (G, M, k etc.) intelligently

It is not the role of this work package to provide a means of converting instrument-related units such as magnitudes, nor to perform coordinate conversions. The former is the domain of the Photometry Data Model to be issued in a separate Data Model effort (See http://www.ivoa.net/cgi-bin/twiki/bin/view/IVOA/PhotometryDataModel for more information and current developments).

Excellent libraries e.g. AST in C (http://www.starlink.ac.uk/~dsb/ast/ast.html) or SLALIB for Java, already exist for the latter. It is our role to meet the needs of these work packages and provide consistency, as well as ensuring that the tools supporting simple conversions (not requiring external or conditional information) are available.

Note on Units is available at: http://www.ivoa.net/forum/dm/att-1495/WD-VOUnits-20090519.pdf.

ACRONYM LIST

1D	One dimension
ADQL	Astronomical Data Query Language
AIDA	Astronomical Infrastructure for Data Access
Aladin	Sky atlas and data discovery tool
ALMA	Atacama Large Millimetre Array
AST	Starlink library for handling World Coordinate Systems
CCD	Charge-Coupled Device
CGI	Common Gateway Interface
CNRS	Centre National de la Recherche Scientifique
D#	Deliverable number
DAL	Data Access Layer
DAL2	Second generation Data Access Layer
DM	Data Model
ESA	European Space Agency
ESA-VO	European Space Agency - Virtual Observatory
ESO	European Southern Observatory
Euro-VO	European Virtual Observatory
EuroVO-AIDA	European Virtual Observatory - Astronomical Infrastructure for Data Access (FP7 Project)
FITS	Flexible Image Transport System
FOV	Field of View
HI	Hydrogen atom
HTTP	Hypertext Transfer Protocol
ID	Identification
IFU	Integral Field Unit
INAF	Instituto Nazionale di Astrofisica
INTA	Instituto Nacional de Técnica Aerospacial
IR	InfraRed
IVOA	International Virtual Observatory Alliance
JRA	Joint Research Activity
MERLIN	Multi-Element Radio Linked Interferometer Network
MOS	Multi-Object Spectroscopy
NVO	National Virtual Observatory (US)
PR	Proposed Recommendation
PU	Public

R&D	Research and Development
RDBMS	Relational Data Base Systems
S*AP	Service * Access Protocol
SED	Spectral Energy Distributions
SI	Abbreviation for the International System of units, from the French Système International
SIA	Simple Image Access
SIA2	Second generation Simple Image Access
SIAP	Simple Image Access Protocol
SINFONI	Spectrograph for INtegral Field Observations in the Near Infrared
SLALIB	Starlink library of positional astronomy routines
SQL	Structured query language
SSA	Simple Spectra Access
SSAP	Simple Spectra Access Protocol
STC	Space-Time Coordinates
SVO	Spanish Virtual Observatory
T#	Task number
TAP	Table Access Protocol
UHEI	University of Heidelberg
UK	United Kingdom
UML	Unified Modelling Language
US	United States
VO or VObs	Virtual Observatory
VOQL	Virtual Observatory Query Language
WG	Working Group
WP#	Work Package number
XML	Extensible Mark-up Language
XMM	X-ray Multi-Mirror Mission