# Europlanet-IDIS Data model:<br>a Data Model for a Planetology Virtual Observatory

**Baptiste Cecconi**$^{(1)}$ **and the IDIS-DM-SWG***

***(1) LESIA-Observatoire de Paris [CDPP]***
*11 Av. Marcelin Berthelot, 92190 Meudon, France*
*EMail: baptiste.cecconi@obspm.fr*

## ABSTRACT

IDIS (Integrated and Distributed Information System) is part of the Europlanet project and aims to develop a prototype of a planetology Virtual Observatory. In the frame of its participation to this project, the CDPP (Data Centre for Plasma Physics, based in Toulouse) is developing a new data model to describe the wide variety of data products that can be found in the planetology community, which includes a wide variety of science thematics such as plasma physics, planetary surfaces, interiors, atmospheres or small bodies. This data model is making extensive use of existing standards provided by various groups (IVOA, SPASE...) and its scope is to describe the scientific contents of datasets, in order to be able to locate and retrieve data files corresponding to a given request.

The initial version has been developed to describe plasma data, which can be very heterogeneous (time series, spectra, dynamic spectra, maps…). Two models are tested in collaboration with VO-Paris (Virtual Observatory Paris Data Centre) and other Europlanet IDIS partners in order to take into account characteristics of data from other thematics of planetary sciences.

Keywords: Virtual Observatory, Data Model, Planetology

***IDIS-DM-SWG**: IDIS Data Model Science Working Group.

*Member list*: Jean Aboudarham[(a)], Nicolas André[(b)], Jérôme Berthier[(c)], Natacha Bourrel[(b)], Maria Teresa Capria[(d)], Baptiste Cecconi[(e)], Maria Cristina De Sanctis[(d)], Stéphane Erard[(a)], Michel Gangloff[(b)], Christian Jacquey[(b)], Maxim Khodachenko[(f)], Pierre Le Sidaner[(g)], Cédric Leyrat[(a)], Nicolas Manaud[(h)], Walter Schmidt[(i)], Bernard Schmitt[(j)], Florian Topf[(f)], Frank Trauthan[(k)], Alain Sarkissian[(l)], Sandrine Vinatier[(a)]

*[(a)]LESIA, CNRS, Observatoire de Paris (VO-Paris), Meudon, France; [(b)]IRAP, CNRS, Univ. Paul Sabatier (CDPP), Toulouse, France; [(c)]IMCCE, CNRS, Observatoire de Paris (VO-Paris), Paris, France; [(d)]INAF, IASF, Roma, Italy; [(e)]LESIA, CNRS, Observatoire de Paris (CDPP), Meudon, France; [(f)]IWF, ÖAW, Graz, Austria; [(g)]DIO, Observatoire de Paris, Paris, France; [(h)]ESAC, Madrid, Spain; [(i)]FMI, Finland; [(j)]IPAG, CNRS, Univ. Joseph Fourier, Grenoble, France; [(k)]DLR, Berlin, Germany; [(l)]LATMOS, CNRS, UVSQ, Paris, France.*

# INTRODUCTION

EuroPlaNet (EPN) is a four-year project supported by the European Union under the Seventh Framework Programme (FP7). This project coordinates a series of research and service activities linked to planetology. IDIS (Integrated and Distributed Information System) is part of this project and aims at defining the grounds of a planetology Virtual Observatory (VO). IDIS is thus a collaborative effort involving scientists from France, Austria, Italy, Spain, Finland and Germany. In the frame of this Research and Development (R&D) activity, several paths are explored. Data Models (DM) and access protocols from various existing groups and projects (such as the International Planetary Data Alliance

(IPDA), the Planetary Data System (PDS), the International Virtual Observatory Alliance (IVOA), etc.) have been studied.

In order to conduct this work, a working group was formed with scientists from the five EPN-IDIS science nodes: Interior and Surface, Atmosphere, Plasma, Small Bodies and Dust, and Planetary Dynamics. These scientists (researchers and engineers) are all involved in data handling and related services in their respective science disciplines. The major challenge of this project is to set the bases of an inter-disciplinary VO. Indeed, there is a large variety of data and instruments covered by the science nodes: remote imaging of atmosphere or surfaces, in-situ measurement of particle distributions or magnetic field, subsurface sounding of icy layers, radio source monitoring, ephemeris of small bodies… This required a strong effort on defining terms and concepts, in order to facilitate exchanges between science communities.

The working group objectives were separated in three tasks: (i) List the various data types and propose a set of query criteria that are relevant for data search in each discipline of EPN-IDIS; (ii) Discuss how to define the PDAP (Planetary Data Access Protocol, a protocol developed by IPDA) extensions that are to be proposed to the IPDA; this second task should be a natural outcome of the first task; and (iii) Propose, define and/or select metadata dictionaries for the IDIS Data Model (i.e. select the standard sources for the possible values describing the data).

We have followed two paths during our study. First we have made an inventory phase to try to identify the set of required metadata for planetology data. Following this study, we have drawn a prototypal DM, keeping in mind that we would link it with PDAP. We have worked with IPDA after an assessment study of PDAP. In parallel, we are also studying the IVOA Observation Core DM (ObsCore), and its related protocol ObsTAP (Observation Table Access Protocol).

The selected DM and protocols will be used for two main purposes. The first one is to share new resources (in case no pre-existing DM/Protocol are already implemented). This is usually the case for high level datasets). The second is to enable basic search in a catalogue (or registry). In each case, two levels of search have been identified: the dataset level search and the granule level search. In the first case, the user looks for datasets that are relevant to its search criteria, while in the second case, the user retrieves the granules (usually the files, see below) corresponding to its search criteria.

# DEFINITIONS

When dealing with conceptual organization of data, it is necessary to clearly define the meaning of the terms used. We propose here definitions for the terms used when describing the organization of a dataset and its granularity. This series of definition is not universal, and other groups working on data archiving and VO have various definitions of the same concepts.

## Resource and Service

We follow here the definitions that can be found in IVOA Documentation[1]:

- A **resource** is a general term referring to a VO element that can be described in terms of who curates or maintains it and which can be given a name and a unique identifier.
- A **service** is any VO resource that can be invoked by the user to perform some action on their behalf.
- A **query service** supports a query/response protocol.

We propose to describe data resources and query services in the field of planetology.

## Data Organization

We define four levels of description for a data resource, from top to bottom: dataset, granule, parameter and record. We also illustrate the definitions with a typical dataset organized as follows: a series of text files containing tables with four columns (time of measurement, measurement A, measurement B, and instrument mode) and in which each line is a record. In this example, the data provider is distributing the data files (either as a whole, or individually). The four levels are defined as follows:

- A **dataset** is a series of data with homogeneous content. This means that all the data records of the dataset must have the same structure. This is the case in our example.
- A **granule** is a homogeneous group of records in a dataset. This definition is further explained in the next section. In our example, each data file is a granule.
- A **parameter** is a series of data identified with a specific quantity. We can define two types of parameters: physical parameters, which are usually the measurement, or the data; and support parameters, which describe the experimental parameters (time, instrument mode…). In our example, there are two physical parameters (columns 2 and 3) and two support parameters (columns 1 and 4).
- A **record** is an individual set of values that cannot be split without loosing the homogeneity property of a dataset. In our case, a record is a line of a file.

It has been argued in our working group that several datasets may be grouped together for archiving purposes (as it is done at the PDS, for instance). This is still under study. Two options are currently envisaged: (i) datasets could be gathered into data collections (for instance, with datasets of different processing level from the same instrument); or (ii) a dataset could contain several parameter sets, in which case the homogeneity property at the dataset level is lost, but is transferred to the parameter set level. This latter case is closer to the PDS data organization.

## Data Granularity

An important aspect is the data granularity, i.e. the size of the data chunks that are made available by the data provider. It is up to the data provider to choose the granularity of his data. Hence, we can identify two extreme cases: (i) the data provider is distributing the full dataset as a whole; (ii) or the data provider is distributing each individual record of his dataset independently. In the first case, the granule (i.e. the minimal portion of data distributed) is the full dataset, while in the second case, each record is a granule. Usually, a Granule is a file (or a series of files, like the 'label' and 'data' files in PDS) containing a series of data grouped over an identified time span.

## DATA MODELS

After studying various DM from other groups, we decided to conduct two studies in parallel. First we have built a genuine DM (IDIS-DM) using bricks and concepts from other DM. Second we are testing the Observation Core DM from IVOA. We present these two approaches here.

## IDIS Data Model

One of the goals of the IDIS DM working group was to identify all the metadata relevant to the data used by our community, in order to build a unified DM for the various datasets of Europlanet science nodes. The IDIS DM should thus be very generic, so that it can cope with such a thematic diversity. This DM provides a semantic description of the data (description of the dataset content), contrarily to a syntactic description (description of the physical organization of the data).

The metadata have been organized in three groups: resource metadata, dataset metadata and parameter metadata (each of these metadata group is described below). A decision has also been made that the resource and dataset metadata groups are compulsory for inclusion into IDIS.

At the time of writing we only studied observational data, but other types of data such as simulation results are to be included in a way that still has to be defined. Furthermore, the description of simulation outputs in a standard way is the scope of another FP7 project called IMPEx[2]. We are working with this group (some of the working group members are also involved in that project) in order to prepare the inclusion of this kind of data in IDIS. We thus deal here mostly with observational data description, with some hints on how we think we will describe other types of data.

## Resource Metadata

The resource metadata are composed of identity metadata (name and identifier), curation metadata (owner, rights, version, publisher), access metadata (access protocol, access URL) and a description (free text field allowing to describe the resource in more details). Dublin Core metadata elements are used here.

## Dataset Metadata

The dataset metadata describe the data provenance, which includes data source type, observation device description, and target description.

The data source type is qualifying the source of the dataset. The allowed values are 'observational', 'laboratory', 'simulation' or 'mixed' (in case of high level dataset built from observational and simulation datasets, for instance). The data source type drives the way the observation device and target information are described.

The observation device description depends on the observatory type: spacecraft, ground based or laboratory experiment. For each of these observatory types, we propose a set of three required keywords: Facility Name, Instrument Name and Instrument Type. The Facility Name stands for the mission name, the observatory name and the laboratory name for spacecraft, ground based and laboratory experiment, respectively. The values associated to these keywords should be extracted from standard lists, such as the IAU observatory code list for ground-based observatories. Such lists do not necessarily exist at this time, but a process to create and maintain them should be initiated. In case of modeled data, the observation device should be the model used. This part of the description has not been studied yet. Among the ideas that have been proposed, two alternatives can be outlined: (i) when modeled data try to reproduce observation of natural targets by existing observers, then, the observation description should state which observatory is simulated; while (ii) in a more general case, modeled data do not mimic an existing observatory, then a fake observatory should be defined including enough description to make the dataset usable. This second case must be studied in more details in the future.

The target description is rather easy for observations of natural objects. The observation target is described in terms of target name and/or target type and location on the target when applicable and/or available. The preliminary list of target types is the following: Asteroid, Comet, Dust, Dwarf Planet, Galaxy, Globular Cluster, Meteorite, Meteoroid, Meteoroid Stream, Nebula, Open Cluster, Planet, Planetary Nebula, Planetary System, Plasma Cloud, Ring, Satellite, Small Body, Spacecraft, Star, Star Cluster, Terrestrial Sample, Trans-Neptunian Object. For laboratory experiments and for modeled data, the target description has not been fully studied yet.

## Parameter Metadata

The parameter metadata characterize the physical or support parameters in terms of physical quantity (what?), coverage (where, when?), processing level (how?) if applicable, etc.

The parameter type indicates if the parameter is a physical or a support parameter.

We propose to define the physical quantity by several means: (i) a UCD[3] (Unified Content Descriptor) is used to roughly define the physical quantity; (ii) references to metadata dictionaries (recommended); and (iii) full text description (only if necessary). In order to describe planetology data, the current UCD list[4] needs to be updated, especially for plasma data. The IDIS working group plans to formulate a series of recommendation in that sense soon.

The coverage metadata include temporal, spatial, spectral and any other type of axes relevant for the parameter. Each axis is defined as with a name, a unit (inspired by the unit description that can be found in the Single Spectral Lines Data Model[5] of IVOA), a reference frame (if necessary) and a support (minimum and maximum values).

The processing level is usually the calibration level of the parameter. At the moment, the identified allowed values: raw, un-calibrated, partially calibrated, calibrated, or derived. The two first values correspond respectively to telemetry data and decompressed telemetry data, while the latest is used for higher-level parameters. We also describe the way the physical parameter has been measured, specifying whether it is a remote or in-situ measurement, or whether it is a passive or active measurement.

It is clear that not all these metadata apply to support parameters, and thus only a subset of them should be used in this case.

## Example

Several examples of dataset descriptors have been built in order to test the various concepts of this DM. Among them, many plasma data sets were tested, as they are usually more complex to describe than images or maps.

We will present here one example showing how critical is the choice of the granularity. The data that we are presenting here have been published by Vinatier et al[6]. This dataset contains vertical abundance profiles in Titan's atmosphere, as derived from an inversion algorithm applied on Cassini/CIRS (Composite InfraRed Spectrometer) spectra observed at Titan's limb. The physical data is composed of a series of text files, each of them containing an abundance profile for a given chemical specie at a given location on Titan (18 limb observations were processed). The inversion provides the vertical abundance for 10 species ($C_4H_2$, $C_6H_6$, $C_2H_6$, $C_2H_2$, $C_2H_4$, $CH_3C_2H$, $C_3H_8$, HCN, $HC_3N$ and $CO_2$). Hence, for each of the 18 locations on Titan's surface, 10 files are available. Considering the dataset homogeneity property, the data provider has two choices: (i) if chemical specie profiles are considered as different parameter, each group of 10 files for a given location are then considered as granules, and they must be distributed together; whereas (ii) if the chemical specie is considered as an axis (which is not possible yet in the DM, but should be included eventually), then each individual file can be distributed independently. In the latter case, an effort should be done in the axes and coverage definition on the DM side, but the description and distribution of such a dataset would be simplified.

Implementation of IDIS DM remains to be done.

# IVOA Observation Core Data Model

The Observation Core[7] DM (ObsCoreDM) has been initially developed to describe astronomical data. We are trying here to use this DM to describe planetary data. This DM has been especially designed for data discovery in astronomy. Hence, it has a very similar goal than our project. The IVOA approach is making use of existing standards that are already used in the astronomy field. An obvious advantage is that data providers that already share their data with IVOA standards (e.g. in IVOA registries) will not have to implement a new DM and a protocol. It would help the integration of existing telescopic planetary images, for instance.

The study of this DM is still at a preliminary stage, but up to now, it seems that the resource and dataset metadata groups defined in the previous section could be handled easily with ObsCoreDM, with very few extensions. As seen on Figure 1, ObsCoreDM is using the Provenance DM (ProvDM) which includes what is described in the observation device description of IDIS DM, except for the instrument type keyword. It also describes the target, but only providing a target name, where we would like to include a target type keyword and the location on the observed body if necessary. The IDIS DM resource metadata are split into the DataID, Curation and Access elements.

Extensive tests and implementation are currently conducted.

# Other Data Models

Other data models have been studied and are actually already used for specific data product within IDIS. Three examples are given here: Space Physics Archive Search and Extract (SPASE), laboratory Solid Spectroscopy and the Open Geospatial Consortium.

## Space Physics Archive Search and Extract (SPASE)

As stated in their webpage[8], the Space Physics Archive Search and Extract (SPASE) effort is a heliophysics community-based project with the goals of (i) facilitating data search and retrieval across the

Space and Solar Physics data environment with a common metadata language; (ii) Defining and maintaining a standard Data Model for Space and Solar Physics interoperability, especially within the heliophysics Data Environment; (iii) Using the Data Model to create data set descriptions for all important heliophysics data sets; (iv) Providing tools and services to assist SPASE data set description creators as well as the researchers/users; and (v) Working with other groups for other heliophysics data management and services coordination as needed.

The Space Physics Archive Search and Extract (SPASE) effort is implemented by the SPASE Consortium, which is composed of representatives of the international heliophysics data community. The SPASE Working Group is currently the only international group supporting global data management for Solar and Space Physics.

Several databases sharing space physics data (plasma physics data) are already implementing this DM and the associated protocol.

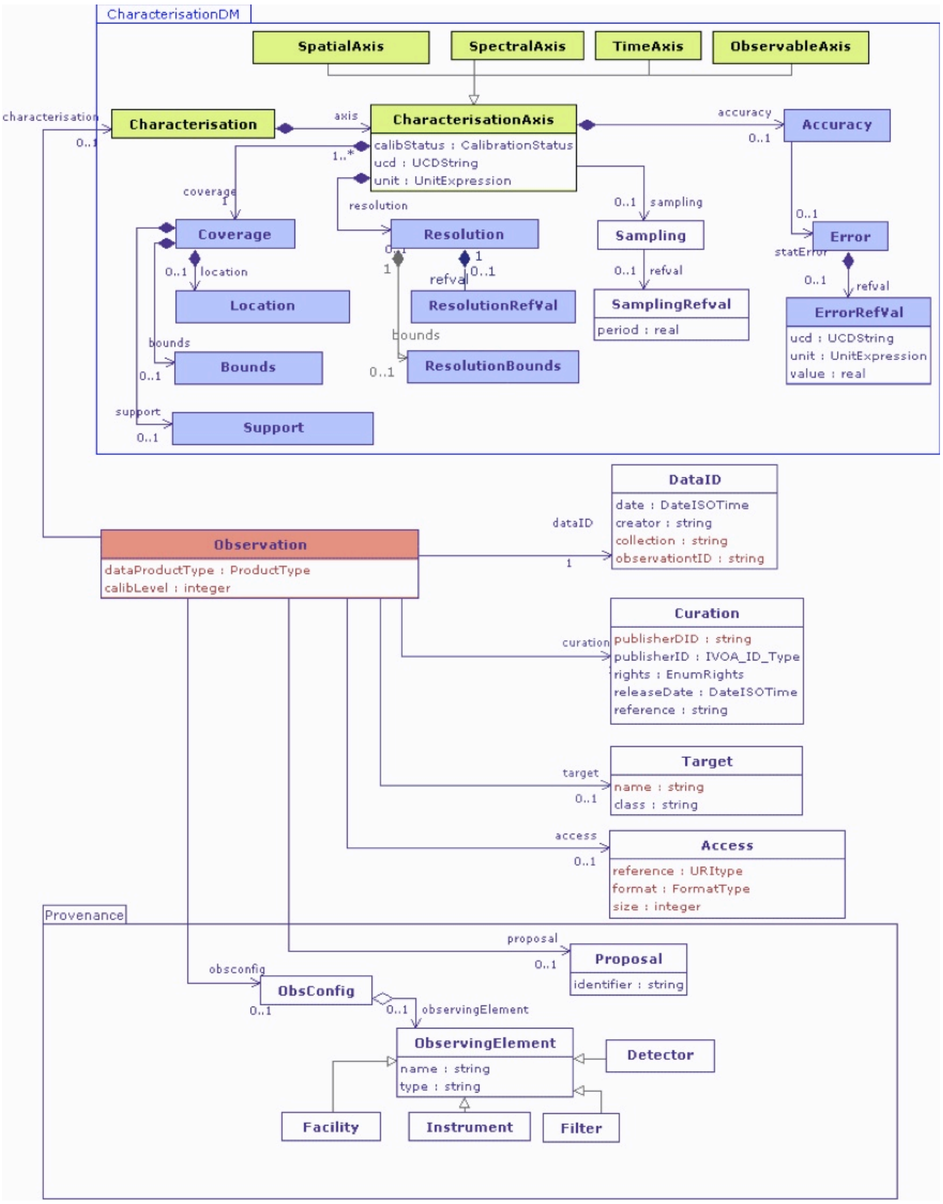

Figure 1: Depicted here are the classes used to organize observational metadata. Classes may be linked either via association or aggregation. The minimal set of necessary attributes for data discovery is shown in brown. (extracted form ObsCoreDM IVOA document[7])

### Solid Spectroscopy DM (SSDM)

SSDM is a DM developed in the frame of GhoSST[9]. It aims at describing two sets of data: (i) experimental data on "spectroscopy of solids" containing a set of sub-databases corresponding to different spectroscopic techniques (IR transmission, vis-IR reflectance, Raman, Fluorescence, IR microscopy) and a "band list" sub-database; and (ii) data on the "physical properties of ices and molecular solids" based on bibliographical reviews and critical analyses of published data (measurements, theoretical calculations…) completed by their own measurements and computations. This group is also developing tools to explore such databases.

## Open Geospatial Consortium (OGC)

This international consortium is a non-profit, international, voluntary consensus standards organization[10]. They are developing standards for geographic content and services, sensor webs, and location services on the surface of Earth. Some planetology groups (especially, scientists from the surface and interior node) are using standard descriptions from OGC data models to describe features on the surface of planets.

# PROTOCOLS

In order to implement a VO prototype, the selection of an exchange protocol is needed. In parallel to the two DM that are studied, two protocols are also under study: Planetary Data Access Protocol[11] (PDAP) developed by the IPDA, and Observation Table Access Protocol (ObsTAP) developed together with ObsCoreDM by IVOA. As we have seen, the two DM propose a very similar level of description, allowing us to implement both protocols, whatever DM is selected. The advantage of their similarities facilitates the interaction with existing PDAP compliant database such as the PDS, the PSA (Planetary Science Archive at ESA), or the JAXA (Japanese Space Agency) data archive, as well as data services and resources already implementing IVOA standards.

## Planetary Data Access Protocol (PDAP) from IPDA

PDAP is born from an effort of worldwide space agencies to come up with some standards for data exchanges. It is based on PDS (version 3) keywords and data dictionaries, but it now has its own namespace. This means that it will evolve independently from PDS. There is an obvious interest to keep track of PDS maturation, but PDAP will be able to follow other DM developments.

PDAP is organized around a PDAP-core protocol, which include basic tools to access datasets and products. In PDS terminology, a dataset corresponds to a data collection described above and a product to a granule. PDAP-core also includes a “map-projected-product” for products that are projected on the surface of a planetary object. The PDAP-core will be enriched using extensions, which will extend the capabilities of PDAP services. PDAP-Core and PDAP-Extensions will be defined and specified in separated documents. Several extensions are already foreseen: Fly-by product (which describes the observations conditions in case there is no planetologic mapping system associated with the observed object, such as comets or asteroids), Spectrum product (for spectral observations), Time-Series product…

The IPDA will release soon (early 2012) the first version of PDAP. At this time, we will begin to test its implementation together with the IDIS DM.

## Observation Table Access Protocol (ObsTAP) from IVOA

ObsTAP is based on the Table Access Protocol (TAP)[12], which is widely used in IVOA. TAP defines a service protocol for accessing general table data. ObsTAP is a version of TAP adapted to ObsCoreDM. The implementation of this protocol is under study in our group.

# CONCLUSION

IDIS aims at prototyping a planetary VO, allowing to discover datasets of interest for a given request in terms of target, time interval, or instrumentation. We have studied two data models (a genuine IDIS DM and the ObsCoreDM from IVOA), as well as two data access protocols (PDAP and ObsTAP). After the DM selection, we plan to implement soon a prototype using one or several protocols. Although this implies a bigger technical effort, the advantage of eventually implementing the two protocols brings interoperability with both IPDA databases and IVOA services. Finally, our ideal IDIS VO client will also have the capability of accessing data shared with other pre-existing protocols and DM such as SPASE, SSDM or OGC.

## REFERENCES

[1] - Resource Metadata for the Virtual Observatory, Version 1.12, march **2007**. http://www.ivoa.net/Documents/REC/ResMetadata/RM-20070302.html

[2] - IMPEx project web page: http://impex-fp7.oeaw.ac.at/overview.html

[3] - An IVOA Standard for Unified Content Descriptors, Version 1.10, August **2005**. http://www.ivoa.net/Documents/REC/UCD/UCD-20050812.html

[4] - UCD website at CDS (Strasbourg, France): http://cdsweb.u-strasbg.fr/UCD/

[5] - Section 3.2 of Simple Spectral Lines Data Model, Version 1.0, December **2010**. http://www.ivoa.net/Documents/SSLDM/

[6] - Vinatier, S., et al. (**2007**). *Vertical abundance profiles of hydrocarbons in Titan's atmosphere at 15° S and 80° N retrieved from Cassini/CIRS spectra*. Icarus, **188**(1), 120–138. doi:10.1016/j.icarus.2006.10.031

[7] - Observation Data Model Core Components and its Implementation in the Table Access Protocol, Version 1.0, July **2011**. http://www.ivoa.net/Documents/ObsCore/index.html

[8] - Space Physics Archive Search and Extract (SPASE) project web page: http://www.spase-group.org/

[9] - Grenoble Astrophysics and Planetology Solid Spectroscopy and Thermodynamics (GhoSST) project web page: http://ghosst.obs.ujf-grenoble.fr/

[10] - Open Geospatial Consortium website: http://www.opengeospatial.org/

[11] - Planetary Data Access Protocol specification at IPDA website: http://planetarydata.org/projects/inactive-projects/data-access/documents/pdap-versions

[12] - Table Access Protocol, Version 1.0, March **2010**. http://www.ivoa.net/Documents/TAP/

The EuroPlaNet RI project is funded by the European Commission under the 7th Framework Program, grant 228319 "Capacities Specific Programme".