Europlanet-RI
Interoperable Data Access
Data Organization Conventions

Version 0.0α1
EPN/JRA4-IDIS/Task 2

This version:

http://typhon.obspm.fr/idis/docs/IDIS-DataOrganization-V0.0alpha0.pdf

Latest version:

http://typhon.obspm.fr/idis/docs/IDIS-DataOrganization-latest.pdf

Previous version(s):

http://typhon.obspm.fr/idis/data-model.html

Editors:

Baptiste Cecconi

Authors:

Baptiste Cecconi, Jérôme Berthier, Michel Gangloff, Nataliya Bourrel, Pierre Le Sidaner, Stéphane Érard, Christian Jacquey, Etienne Pallier, Nicolas André, Nicolas Lormant, Florian Topf

Abstract

This document defines the data organization convention adopted for dataset description. We propose a hierarchical data organization: Dataset, Granule, Record and Parameter. We also describe resources and services following IVOA recommendations.

Status of this document

This document is a draft and should be approved by the IDIS-Data-Model science working group.

Acknowledgements

This work has been funded by EC (Grant agreement n°228319). This document has been edited from the ObsCoreDM document from IVOA.

Contents

List of Acronyms 2
Introduction 3
Resource and Service 3
Data Organization 3
Data Granularity 4
Discussion 4

List of Acronyms

DM	Data Model
IDIS-DM	IDIS Data Model
IDIS-TAP	TAP interface to IDIS-DM
IVOA	International Virtual Observatory Alliance
PDAP	Planetary Data Access Protocol
PDS	NASA Planetary Data Science Archive
TAP	Table Access Protocol
VO	Virtual Observatory

Introduction

When dealing with conceptual organization of data, it is necessary to clearly define the meaning of the terms. 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 here to describe data resources (using IDIS-DM) and query services (using IDIS-TAP or PDAP) in the field of planetary sciences.

Data Organization

We define four levels of description for a data resource, from top to bottom: **dataset**, **granule**, **parameter** and **record**. We 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.

[1] http://www.ivoa.net/Documents/REC/ResMetadata/RM-20070302.html

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.

Discussion

The main discussion point is on the "single granule type per dataset" paradigm. Datasets are not defined in such a way in PDS archives for instance. In our proposed Data Organization Convention, different types of granules (for instance, for various calibration levels) are grouped into so many different datasets. At the moment, our position is to have Dataset collections, which are grouping the datasets as defined here.