Data Request XML format

Martin Juckes, November 14th, 2017.

1 Executive Summary

The Data Request defines the diagnostics which are requested from modelling centres participating in the sixth phase of the Climate Model Inter-comparison Project (CMIP6).

The Data Request is presented as two XML files: a configuration file and the content. Each file has an associated XSD schema. The XSD schema for the content file is generated automatically from the configuration file. For many users it will be more convenient to deal with the python interface or web and spreadsheet versions of the request, which will be described in a separate document. The transformation to an XML format from the traditional spreadsheet format is designed to deal with a number of issues associated with growing complexity and a need to support automation driven by the scale of the request. In order to preserve continuity, many of the records in the XML files will have a direct relation to spreadsheet rows in the traditional format.

A separate document describes a simple python API for the data request.

The variables are now also listed in a spreadsheet of MIP tables:
proj.badc.rl.ac.uk/svn/exarch/CMIP6dreq/tags/latest/dreqPy/docs/CMIP6_MIP_tables.xlsx

In release 01.beta.27 a supplement was added, containing records on quality control and (under development) information about physical relationships between variables. The supplement is presented as a separate file in order to simplify the version management of the main request document.

2 Objectives

The broad objectives of the data request are:

(1) Define variables, together with technical information required for generation of output files;

(2) Define collections of variables, from specified experiments, which are needed for or relevant to specific scientific objectives;

3 Files

The framework schema:

http://proj.badc.rl.ac.uk/svn/exarch/CMIP6dreq/tags/latest/dreqPy/docs/vocabFrameworkSchema_01beta.xsd

Configuration file:

http://proj.badc.rl.ac.uk/svn/exarch/CMIP6dreq/tags/latest/dreqPy/docs/dreq2Defn.xml

Data request schema:

http://proj.badc.rl.ac.uk/svn/exarch/CMIP6dreq/tags/latest/dreqPy/docs/dreq2Schema.xsd

Data request XML:

http://proj.badc.rl.ac.uk/svn/exarch/CMIP6dreq/tags/latest/dreqPy/docs/dreq.xml

Supplement schema:

http://proj.badc.rl.ac.uk/svn/exarch/CMIP6dreq/tags/latest/dreqPy/docs/dreqSuppSchema.xsd

Supplement:

http://proj.badc.rl.ac.uk/svn/exarch/CMIP6dreq/tags/latest/dreqPy/docs/dreqSupp.xml

4 Overview

Configuration file

The XML Data Request is presented as a configuration file and a content file.

The configuration file contains three types of information:

(1) Layout information which is used to generate the content schema;

(2) Comments on the purpose and intent of attributes;

(3) Technical labels to facilitate automated navigation of the contents.

If users wish to exploit the XML files directly it is recommended that they make use of the configuration file, as the information types (2) and (3) are not embedded in the content file.

Each section of the document is defined by a “table” element with the following attributes:

- label (e.g. 'var'): a name for a section of the content – will be used as the XML element name;
- title (e.g. 'MIP variable”): a longer, human readable string;
- id: an opaque name;
- itemLabelMode: specifies whether the “label” attribute of records in this section should permit use of '-';
- level: an integer, designed to assisted automated processing by giving an indication of the structure of the request;
- maxOccurs: maximum number of times the section is allowed;
- labUnique [Yes|No]: set to yes if label values for records are unique within each section.

Within each section there are definitions for attributes of items. Each item attribute is defined using the following configuration attributes:

- label: this will be the attribute name;
- title: a longer string explaining usage;
- class: the class supports automation. e.g. attributes which refer to another record in the document will have the class set to “internalLink”;
- type: the xsd content type (e.g. “XS:STRING”);
- techNote: to support automation. e.g. if class is “internalLink”, this attribute should be set to the name of the intended section:
- required: indicates whether the attribute is required;[1]
- usage: notes on the usage of the attribute.[1]

Since 01.beta.33 all properties (title, valid_max, valid_min, etc) in all sections have been given a

1 New in 01.beta.17

valid title to aid legibility of the document: the titles of the attributes can be used, for instance, to provide help notes in a web display of the data request.

In addition to the standard XSD content types “string”, “boolean”, “integer”, “duration” and “float”, the following types are defined:

- st__integerList: a list of integers;[1]
- st__integerListMonInc: a monotonic increasing list of integers (monotonicity is not checked by the XSD schema, but is verified by the python API);[1]
- st__floatList: a list of integers;[1]
- st__stringList: a list of words;[2]
- st__attLabel_def: a string composed of characters “a” to “z”, “A” to “Z”, “0” to “9” and “-”;
- st__attLabel_und: a string composed of characters “a” to “z”, “A” to “Z”, “0” to “9” and “_”;
- st__attLabel_an: a string composed of characters “a” to “z”, “A” to “Z”, “0” to “9”;
- st__uid: a-zA-Z0-9:_.+-

The following table summarises the specifications of the core attributes:

Table 1: Core attributes

label	**title**	**description**	**usage**
label	Record Label	A single word, with restricted character set	A short mnemonic word which is potentially meaningful but also concise and suitable for use in a programming environment
uid	Record Identifier	Unique identifier	Must be unique in the data request. For well known concepts this may be related to the label, but for items such as simple links between concepts an a random string will be used.
title	Record Title	A few words describing the object	A short phrase, suitable for use as a section heading
description	Record Description	An extended description of the object/concept.	
useClass	Record Class	The class: value should be from a defined vocabulary. All records in the schema definition section must have class set to "__core__".	The useClass declared for an attribute can affect its interpretation in the Python package. For example, attributes labelled as “useClass=internalLink” should refer to another data request record.
type	Record Type	The type specifies the	

2 New in 01.beta.19

		XSD value type constraint, e.g. xs:string.	
techNote	Technical Note	Additional technical information which can be used to specify additional properties.	
superclass	Superclass	States what class the property is derived from	
id	Alternative identifier	Alternative identifier	For sections, the id provides a short alias for the section label.
itemLabelMode	Item Label Mode	Item Label Mode	
level	Level	Level	Redundant
maxOccurs	Maximum number of permissible occurrences of this section	Maximum number of permissible occurrences of this section	Used in defining sections. In the CMIP6 Data Request each section only occurs once.
labUnique	Set true if label of each record is unique within section	Set true if label of each record is unique within section	Used in defining sections.
usage	Usage notes	Notes on the usage of the predicate/concept defined by this node	

The above attributes provide the framework for detailed description of data request attributes and diagnostics.

Content file (dreq.xml)

The content file contains three elements at the top level: “prologue”, “main” and “annex”[3]. The “prologue” contains Dublin Core metadata describing the document and a PAV version attribute holding the document version[4]. The “main” element has the sections specified in the configuration file, and within each section a list of records (“item” elements). Each item element has attributes as specified in the configuration file, a different set of attributes for each section. There are no child elements or text content, all the information is in the defined attributes. Every item, across all sections, will have at least these 3 common attributes which are intended to give basic information about the item, thus enabling standardisation in error tracking:

3 New in 01.beta.16

4 New in 01.beta.29 (purl.org/pav/2.3)

- uid: an identifier which is unique within the document;
- label: a short name, using only the characters a-z, A-Z, 0-9 and '-' (in some sections the '-' is disallowed);
- title: a longer name.

The “annex” element also contains a list of sections with the same structure as in the “main” element. The “annex” has been introduced to allow some flexibility in the version management.

Sections

There are 35 sections in the current document, 6 of which contain information about variables, output format and their priorities. An index to the request sections is available here: http://clipc-services.ceda.ac.uk/dreq/index.html .

The sections, with section numbers, are listed below:

1.1 Model Intercomparison Project [mip]

1.2 MIP Variable [var]

Each MIP variable record defines a MIP variable name, associated with a CF Standard Name.

1.3 CMOR Variable [CMORvar]

Each Output variable record corresponds to a MIP table variable specification. In a change from the August draft, this record does not contain the “priority” attribute: the priority is now set in the “Request Variable” record. The other change is that a collection of attributes specifying dimensions have been moved into the “structure” record, and each “CMOR Variable” record links to one structure record. This will facilitate provision of clear and consistent definitions of output formats.

1.4 Request variable (carrying priority and link to group) [requestVar]

The request variable is now a short record which combines a CMOR variable with a priority and assigns it to a request group. The request variable records define the contents of each

request group.

1.5 Experiments [experiment]

The experiment record contains the key information from the “Experiment” sheet of the request template, including the tier of the experiment, the duration and start/end dates.

1.6 Scientific objectives [objective]

The objectives defined by each MIP can be used to select data requirements.

1.7 Specification of dimensions [grids]

A section for the CMOR dimensions specifies the structure of the axes of the requested diagnostics.

1.8 CF Standard Names [standardname]

The reference list of CF standard names is provided at cfconventions.org, but the definitions of terms used in the data request are copied into this section so that the detailed definitions are easily accessible to data request users.

1.9 Experiment Group [exptgroup]

The experiment group defines a collection of experiments within a MIP which might be part of a collective data request.

2.1 Spatial dimensions [spatialShape]

The spatial shape record contains the spatial dimensions of the field, and also, for convenience, an integer specifying the number of levels if that number is specified. A boolean level flag is set to

“true” if the number of vertical levels is specified.

2.2 Temporal dimension [temporalShape]

The temporal shape record contains the temporal dimensions.

2.3 Dimensions and related information [structure]

The structure record combines specification of dimensions, cell_measures and cell_methods attributes. Spatial and temporal dimensions are specified through links to “spatialshape” and “temporalshape” records.

2.4 MIP tables [miptable]

3.1 Request variable group: a collection of request variables [requestVarGroup]

The request variable groups collect variables.

3.2 Request Item: specifying the number of years for an experiment [requestItem]

The request item links a collection of variables with a specific experiment or group of experiments, and a temporal range for output. The “esid” attribute links to an experiment, and experiment group or a MIP. In the latter case, the request applies to all experiments defined by that MIP. The Request Item includes a “Tier Reset” attribute (“treset”)[5] which can override the Tier assigned to the experiments identified by “esid”. Has an optional link to a time slice[3].

3.3 Request link: linking a set of variables and a set of experiments [requestLink]

The request link records specify some additional information about variable groups, concerning shared output requirements and objectives.

3.4 CMOR Table Sections [tableSection]

3.5 Model configuration options [modelConfig]

3.6 Links a variable to a choice element [varChoiceLinkC]

Presence of a link indicates that there is a choice of different representations for a diagnostic.

3.7 Link between scientific objectives and requests [objectiveLink]

Each objective link record joins one objective to one request link. Some requests are linked to multiple objectives and most objectives are linked to multiple requests.

3.08 Remarks about other items [remarks]

The remarks section contains additional comments about other records. It can be used to add detail without adding to the complexity of the other sections.

3.09 Links a variable to a choice element [varChoiceLinkR]

Indicates that there is a ranked choice of variables, and that only one of the ranked list is required.

3.10 Indicates variables for which a there is a range of potential CMOR Varibles [varChoice]

There are several instances where variables defined in the tables are mutually exclusive options of which only one should be requested. The varChoice section is designed to hold this information, but is not yet complete. Examples are between ocean cell volume on a fixed grid for some models and monthly means for others, or between 6 hourly pressure level data on 8 levels vs. 4 levels for different objectives in the HighResMIP request.

3.11 Time Slices for Output Requests [timeSlice]

Specifies time slices (i.e. subsets of an experiment when data for the full duration of the experiment is not required).

4-5: section omitted for possible later use.

---

5 New in 01.beta.17

6.1 Tags

Tags related to processing requirements associated with some diagnostics to aid automated processing.

6.2 Relations between CMOR variables [varRelations][6]

Provides structured information about the difference between variables of the same name and frequency in different tables. E.g. different masking, temporal mean vs. point, different vertical structure (model levels vs. pressure levels).

6.3 Variable relation link [varRelLnk]

Provides links between CMOR variables and varRelation records.

7.1 Cell Methods [cellMethods]

Quality Control Ranges [in supplement]

Extends the information provided in the valid_min, valid_max, ok_mean_min_abs, ok_mean_max_abs attributes which were present in the CMIP5 CMOR tables. In this section there are also attributes valid_max_status etc which indicate the level of confidence in the suggested limits:

- **robust**: A well characterised limit based on a rigorous constraint (e.g. and area fraction must be between 0 and 1) or on a large ensemble of consistent model results.
- **suggested**: A limit which may not be reliable, but which is based on a range of models or plausible arguments.
- **tentative**: Very limited information – e.g. only one or two models in CMIP5 provided the parameter.

Further discussion is available in a draft document on Quality Control range[7], and web pages presenting a review of CMIP5 ranges shows the information being used to construct the control values[8].

Places, States or Reservoirs [places]

Transfers of Material [transfers]

Units [units]

X.1 Core Attributes [core]

The attributes listed in table 1 above.

X.2 Data Request Attributes [main]

Attributes used to in the content records, such as “units”, “valid_max”. Each record in this section defines one of these attributes, specifying its type and other properties used in the python API.

X.3 Section Attributes [sect]

Defines the attributes which are used to describe each section.

Diagrammatic view of Data Request sections

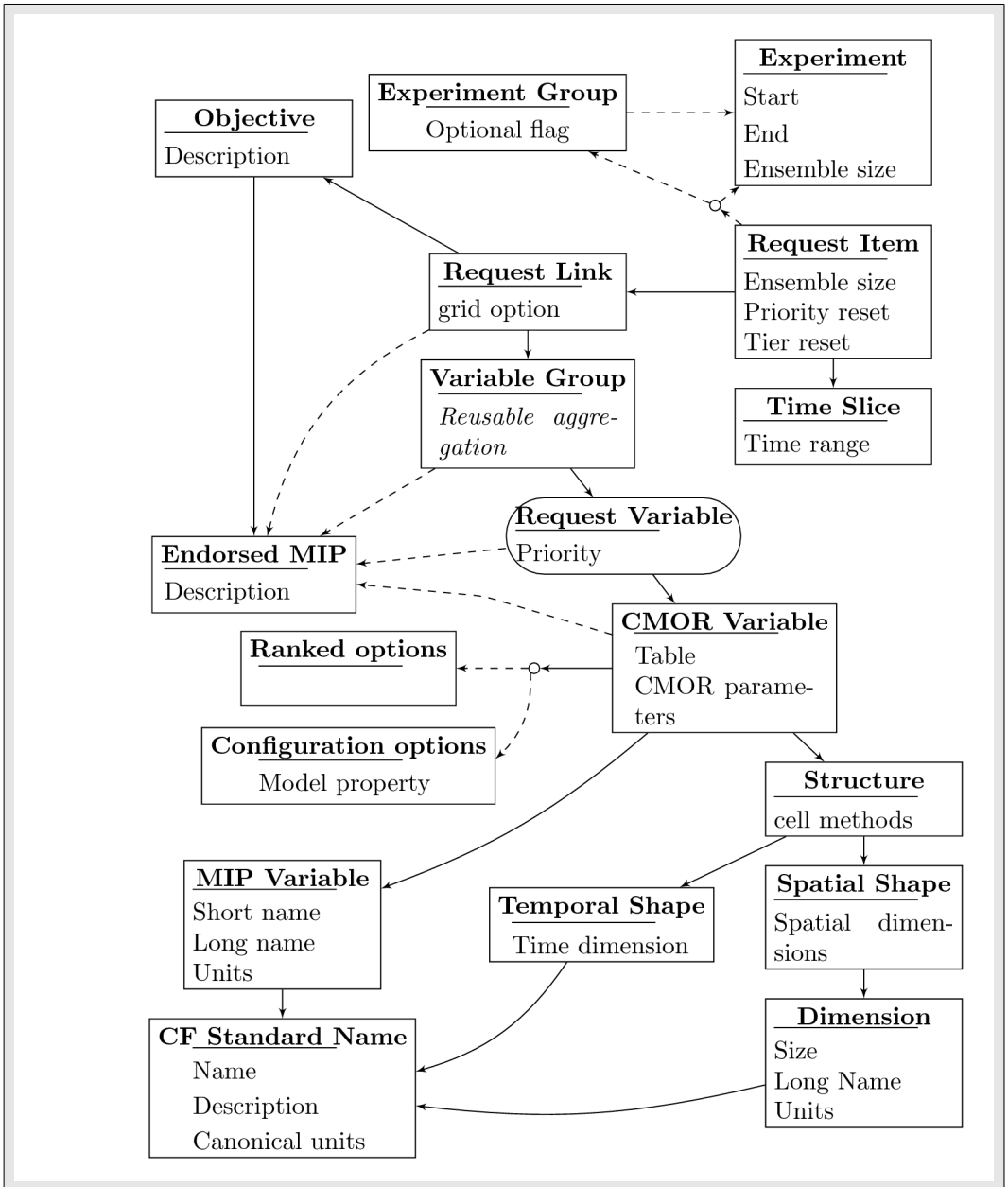
The following diagram illustrates the links between the different sections.

---

6 There appear to be a number of broken links in this area .. the use of these records is under development.

7 https://docs.google.com/document/d/1cvSphy3Hb07t92BJvtqEBM9DMbsOSdENbwLJxw4AmH8/

8 http://clipc-services.ceda.ac.uk/ranges/ or http://w3id.org/cmip6dr/ranges/day_clt.html for a direct link to a single variable.



Caption: Linkage between data request elements. The Endorsed MIPs link to several different records because of the multiple roles they play in creating the request. CMOR variables, for example, link to the MIP which is responsible for defining the variable. The complete list of variables requested by a MIP, which typically includes many variables defined in CMIP5, is obtained by following the the request links associated with that MIP, filtered by experiment(s), priority and objective if you are interested in a selective list. Dashed lines indicate links which are optional or supplementary. Solid lines indicate the primary links needed to decipher the request.

5 Discussion

The layout of the variable definitions has been rationalised into 5 sections: the “MIP variables”

defining the physical parameters, “structure”, “spatialShape” and “temporalShape” defining output configuration and a “CMOR Variable” bringing all these together. The Request Variable table then links CMOR variables together in Request Groups. The request groups give the MIP coordinators the ability to pick and choose precisely the variables needed for each analysis, avoiding requests for unnecessary data. This will result in request groups which contain overlapping data requirements. The use of links back to CMOR variables make it possible to unambiguously determine the union of any set of request groups.

The sections on structure and shape separate out different aspects of the CMOR variable specification and make it possible to ensure that terms are used consistently. The contents if these sections in this draft have been created by scanning the CMOR tables, and there is some duplication (e.g. the cell_measures variable attribute is set for some variables and omitted for others, creating two sets of structure records which are identical except for this distinction. In CMIP6 the cell_measures attribute will always be set).

The link between the request items and the experiment definitions is not fully implemented in this version, but the links through to the variables are. This means it is possible to gain an estimate of the data volumes for each MIP and for combinations of MIPs, but not yet to select specific tiers in a clean way (see dreqPy.pdf for more details). The data volumes given by the current version should be treated with caution. The contents may not fully reflect the intentions of the MIP coordinators, and there may be adjustments to variable priorities.

6 Annex 1: Known issues

The document is still contains some irregularities. The major ones, many of which will be addressed in the next few weeks, are listed below. This list will become more comprehensive in future releases.

……