Data Request XML format

Martin Juckes, September 25th, 2015.

1 Executive Summary

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.

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/01.beta.02/docs/vocabFrameworkSchema_01beta.xsd

Configuration file:

http://proj.badc.rl.ac.uk/svn/exarch/CMIP6dreq/tags/01.beta.02/docs/dreq2Defn.xml

Data request schema:

http://proj.badc.rl.ac.uk/svn/exarch/CMIP6dreq/tags/01.beta.02/docs/dreq2Schema.xsd

Data request XML:

http://proj.badc.rl.ac.uk/svn/exarch/CMIP6dreq/tags/01.beta.02/docs/dreq.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.

Content file (dreq.xml)

The content file contains two elements at the op level: "prologue" and "main". The "prologue" curently contains a short document title, but will be expanded to contain standard document metadata. 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 attributes:

- uuid: 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.

There are 15 sections in the current preliminary document, 6 of which contain information about variables, output format and their priorities.

Sections

1. MIP variables

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

2. CMOR Variables

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 etc 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.

3. Request Variables

The request variable is now a short record which combines a CMOR variable with a priority and assigns it to a request group.

4. 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.

5. Spatial shape

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.

6. Temporal shape

The temporal shape record contains the temporal dimensions.

7. Request variable groups

The request variable groups collect variables.

8. Request link

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

9. Request item

The request item links a collection of variables with a specific experiment or group of experiments, and a temporal range for output. At present the “expt” field is text which is intended to match names of experiments or experiment groups. In future this will be replaced with appropriate record identifiers to support fully automated processing.

10. 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 and end dates.

11. Experiment group

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

12. Objective

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

13. objectiveLink

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

14. 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.

15. 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.

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 (provided that there is no duplication of variable in the CMOR variables section – removal of duplicates will be a priority in the coming weeks).

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.