

Draft Note on Information Flow within the ESG Federation

Bryan Lawrence et al

Introduction

The aim of this document is to outline:

1. Where information entities are expected to flow within the ESG federation,
2. What software is needed to transform information entities, and
3. How information is brought together.

This version of the document may be outright wrong in places. It's a strawman ... please help us correct it.

Key Information Artefacts and their creation.

1. The Data Reference Syntax¹.
2. PCMDI controlled vocabularies (primarily those required by the DRS).
 - Created by Karl Taylor. Status? Where is the official list (e.g. model names, centre names etc). How will it be updated. Where will it live?
3. CMOR Tables:
 - Created by Karl Taylor and Charles Doutriaux. Status nearly complete. Publicly available².
4. NetCDF files (conforming to CMOR conventions).
 - Created by the modelling groups. Exposed via ESG data nodes. Updated by the modelling groups.
5. Directory Structures (conforming to DRS conventions)
 - Created by the modelling groups and laid out on the disk underneath the ESG data nodes. The modelling groups will need to conform to the DRS and the controlled vocabularies.
6. Gridspec (NetCDF) files:
 - To be created by the modelling centre. (Or are they going to be produced from the data files by ESG publisher). Where do these live? In the DRS? One per model component? Are the gridspec files themselves CF compliant? Do we stitch them together using the model/realm name? What is the name of these files. (It appears that realm is the logical place to do this)? Where is the regridding software? How is it connected? What is the catalog showing?
 - CIM documents can't be created from Gridspec
7. ESG Publisher databases

1 CMIP5/AR5 Data Reference Syntax: http://cmip-pcmdi.llnl.gov/cmip5/docs/cmip5_data_reference_syntax_v0-20_clean.pdf

2 See svn repository at <http://www2-pcmdi.llnl.gov/svn/repository/cmor/trunk>

V0.1 Initial Draft (BNL 27/10/09)

- Created by parsing the contents of the directories. Updated by reparsing? Schema defined where? (uses sql-alchemy). Relation to TDS catalogs?
8. Replication Inventories:
 - To be created by ESG publisher and used by proposed replication software to both deliver and verify delivery.No further discussion until the replication procedure is clearer.
 9. Data Node Thredds Catalogs
 - Created by ESG publisher. Updated by ESG publisher.
 - Where and how is the notion of a service endpoint defined?
 10. ESG Gateway databases and Catalogs
 - Configured to receive publication information from ESG data nodes (via TDS catalogs) and other gateways (via OAI).
 - Receives Curator OWL documents via a manual process. Can we automate this. Luca would like to use OAI. Bryan would like to use atom-pub-sub.
 - How does it align model and data?
 - How does it align gridspec, data and models?
 - What is the internal data model?
 - What is the format of the material moved around by OAI between gateways?
 - What does it do about service endpoints?
 11. CMIP5 Questionnaire Internal Structures
 - Primary structure defined in python³
 - Vocabularies defined as mindmaps⁴ and in python⁵
 - Some elements of vocabs need to conform to what is required for stitching data to metadata.
 - Do we need some top-level stuff
 12. Metafor CIM documents describing CMIP5 models, simulations etc.
 - Conforming to CIM V? Defined at ?
 - CIM UML gridspec data model (Ian's UML) is different from the NetCDF gridspec (which has evolved since). Need code to get CIM representation from Gridspec. Do we want to parse uploaded gridspec files and add extra questions? Role of the new GML change request.
 13. ESG Curator OWL files
 - To be produced from the Metafor CIM documents (at least experiment, simulation, component, platform).

3 File models.py at <http://metaforclimate.eu/svn/cmip5q/trunk/cmip5q/cmip5q/protoq>

4 See mindmaps etc at http://metaforclimate.eu/svn/controlled_vocabularies/trunk

5 See XMLinitialise.py at <http://metaforclimate.eu/svn/cmip5q/trunk/cmip5q/cmip5q>

Schematics

This schematic needs to be replaced with something more complete, showing all the artefacts, who is responsible for them, and who is responsible for the transitions.

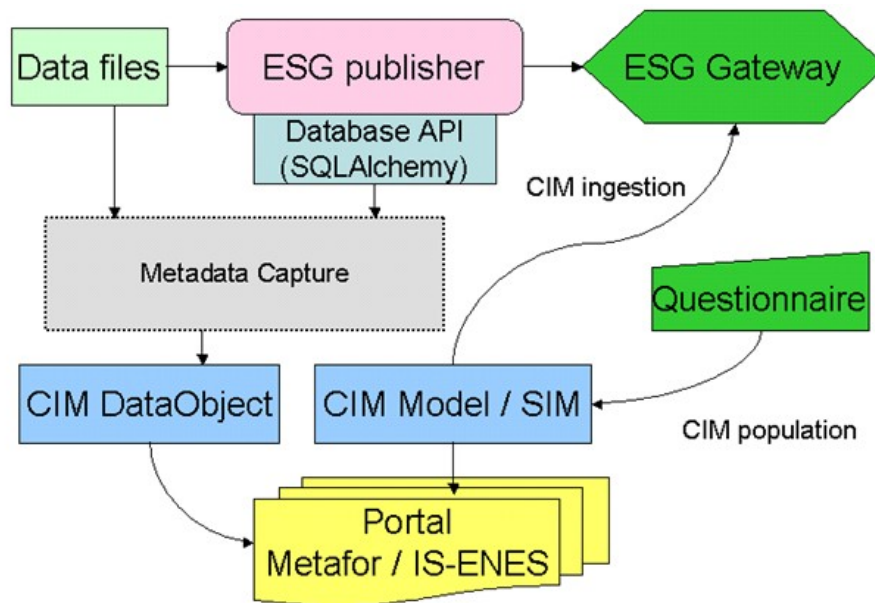


Figure 1: Possible relationship between IS-ENES and ESG portals and their information flow (from Metafor Deliverable 6.1)

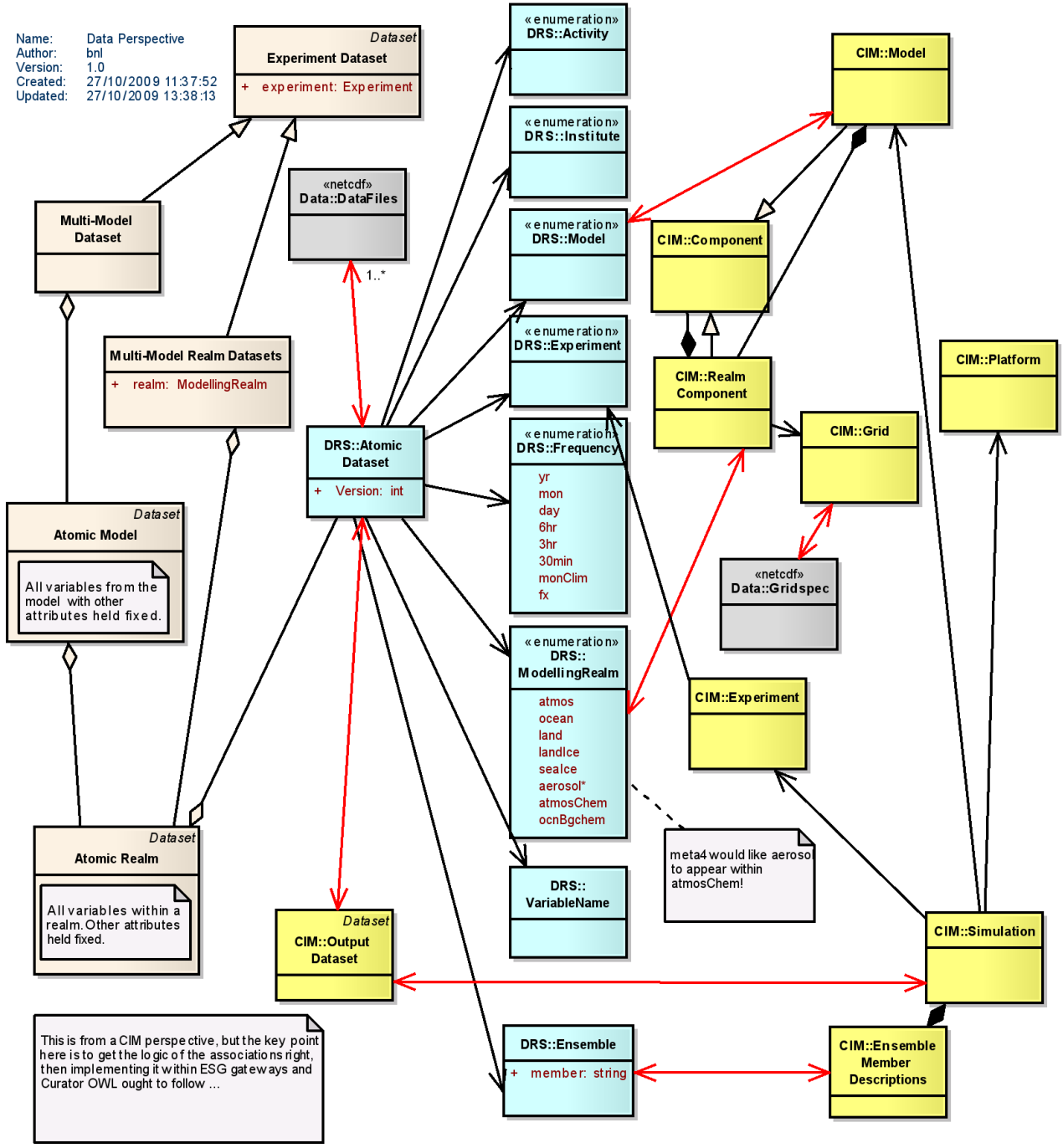
Other issues associated with this schematic are:

1. Where is gridspec?
2. How are the various entity merges accomplished? What are the rules associated with them? Regardless of whether IS-ENES and ESG have different technology portals, the basic merge logic has to be the same.

Which brings us to ...

What is a dataset?

Metafor Perspective?



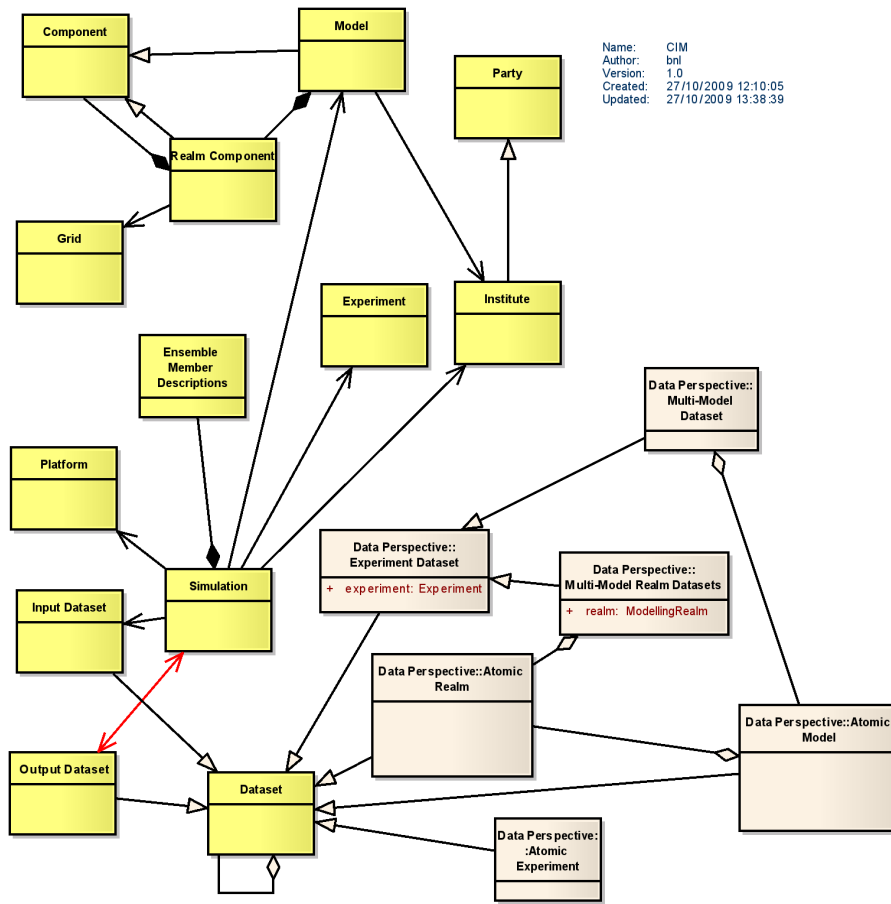


Figure 3: Key CIM entities and their relationships to key dataset concepts.

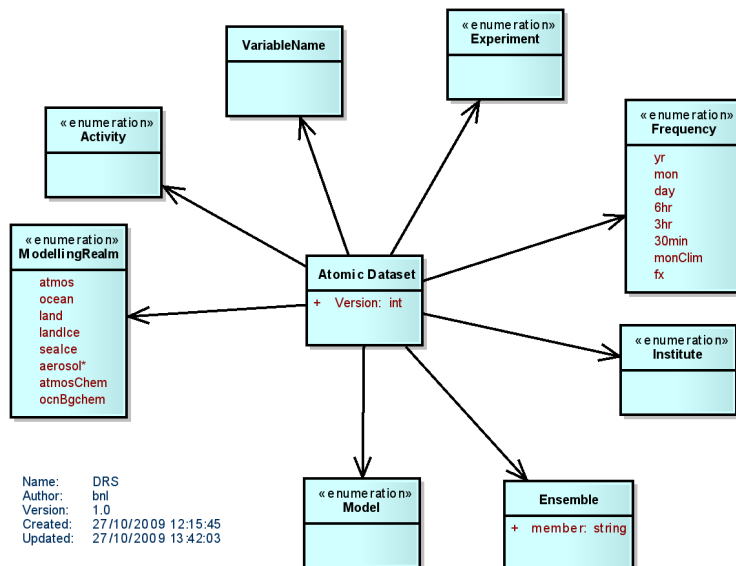


Figure 4: Key DRS entities