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.

Document History

V0.1 Initial draft, 27th of October.
V0.2 Annotated following go-essp-tech telco with Charles, Dean, Karl, Bob, Bryan, Roland and Balaji attending.

Key Information Artefacts and their creation.

1. The Data Reference Syntax[1].
2. PCMDI controlled vocabularies (primarily those required by the DRS).
    - Created by Karl Taylor. Some are in the DRS, some will be in a spreadsheet to appear on the PCMDI site, which will be be the authoritative version. Karl will keep a version number in the document or it's name.
    - The initial spreadsheet is expected in appear by about 18 November (based on a poll currently underway). (Action Karl)
    - A vocab server version of the tables within it will be needed (Action Metafor)
3. CMOR Tables:
    - Created by Karl Taylor and Charles Doutriaux. Status nearly complete. Publicly available[2].
    - A discussion about realms led to a decision for Karl to revisit the arguments and make a proposal/decision and communicate it within a week or so (3rd November, Action Karl).
    - A discussion about atmospheric chemistry led to Karl suggesting he'd check the satus of those tables. (Action: Karl)
    -

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

4. NetCDF files (conforming to CMOR conventions).
    - Created by the modelling groups. Exposed via ESG data nodes. Updated by the modelling groups.
    - Note that netCDF 4.0.1 or higher will be needed. Zlib compression allowed. NOT slib.
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.
    - There are some issues with the DRS that need resolving. In particular, how the DRS should support regrdded fields and other processed outputs. Action: the DRS team to resolve.
6. Gridspec (NetCDF) files:

    (Not all decisions final).

    - Gridspec files for complicated model grids should be produced by the modelling centres with the complicated grids.
    - Gridspec files for simple grids can be produced directly from the archive contents. Action Charles.
    - Gridspec files will appear in the directory structure of the DRS with variable name=gridspec, one per realm, duplicated in every experiment. Action Charles: to email an example URL and gridspec filename.
    - These gridspec files are suitable for regridding, and provide some of the information which one would want to see in the ESG and Metafor catalogues, but not all.
    - The production of CIM XML grids is a Metafor problem, but would be aided by the ues of a gridspec to XML converter. Action: Metafor
    - CIM XML grids to ESG would become a Sylvia problem . Currently they have no way of loading grid descriptions automatically anyway. Action: Bryan to discuss with Sylvia.
    - Questions Unresolved (due to insufficient telco time):
        - Are the gridspec files themselves CF compliant?
        - What is the catalog showing?
        - Gridspec documents are primarily devised for regridding. CIM XML grids are to support **discovery** and **understanding** as well as point to the complete gridspec necessary to carry out **calculations**. There is currently a mismatch between these objectives that needs to be resolved so we can ensure we meet all three objectives.
    - There is going to be a command line tool which users gridspecs to do regridding which exploits the gridspec regridgin being integrated into libcf. Charles will do a python API to libcf (easy_installable?) It will be absorbed into CDAT. Action Charles. A Ferret API will also exploit libcf. Action Roland.
    - There was some discussion of python NCML as well. Bob's done some preliminary work on a rudimentary NCML in CDAT. Bryan observed that his group had a requirement for this and Roberto De Almeida had planned to work on this too.

Action; Bryan to chase up whether we can get a project together to make this happen on a useful timescale.

7. ESG Publisher databases
    - Has configuration file with options described at:
      http://www2-pcmdi.llnl.gov/Members/bdrach/.personal/esg-publisher-configuration
      which defines the relationships with the TDS.
    - Data information created by parsing the contents of the directories, essentially what this accomplishes a replication of netcdf attribute semantics in the database. Schema defined where? (uses sql-alchemy).
    - The relationship between ESG publisher, reparsing, and replication has yet to be determined. Action: Bob et al.
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?
        - Services are described in the TDS associated services.
        - LAS services will be configured into the data node stack, and will appear in the TDS.
        - Local services are not yet supported. The relationship between data entities, and the services exposed for them appears in the TDS, but how this is configured in ESG publisher has yet to be determined. Action Bob
10. ESG Gateway databases and Catalogs

    (None of this was discussed on the 27th , since we didn't have a gateway representative on the telco. We'll try and hijack a Curator call to discuss most of this material. Action Bryan to setup.
    - 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 python[3]

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

- Vocabularies defined as mindmaps[4] and in python[5]
- Some elements of vocabs need to conform to what is required for stitching data to metadata.
- Still some discussion as to the right place for Aeorosol and Atmospheric Chemistry, see point 3 above).
- What support does the questionnaire have for post-processing descriptions. (Answer very little. Action: Bryan to add support in the questionnaire, but that may not be necessary in t he first release).
- Do we need some top-level stuff for grids?Yes, see grids discussion above. Action; Bryan to support to some extent in questionnaire.

12. Metafor CIM documents describing CMIP5 models, simulations etc.
    - Conforming to CIM V? Defined at ?. Action: Metafor

13. ESG Curator OWL files

    Not discussed on the 27th. To be discussed with #10 on a Curator call. Action Bryan.

    - To be produced from the Metafor CIM documents (at least experiment, simulation, component, platform). Probably Grid too.

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

Basic schematic of the key components which need to exist in the information realm. A more complete diagram would expand on the relationships inside the data box and their influence on data flow.

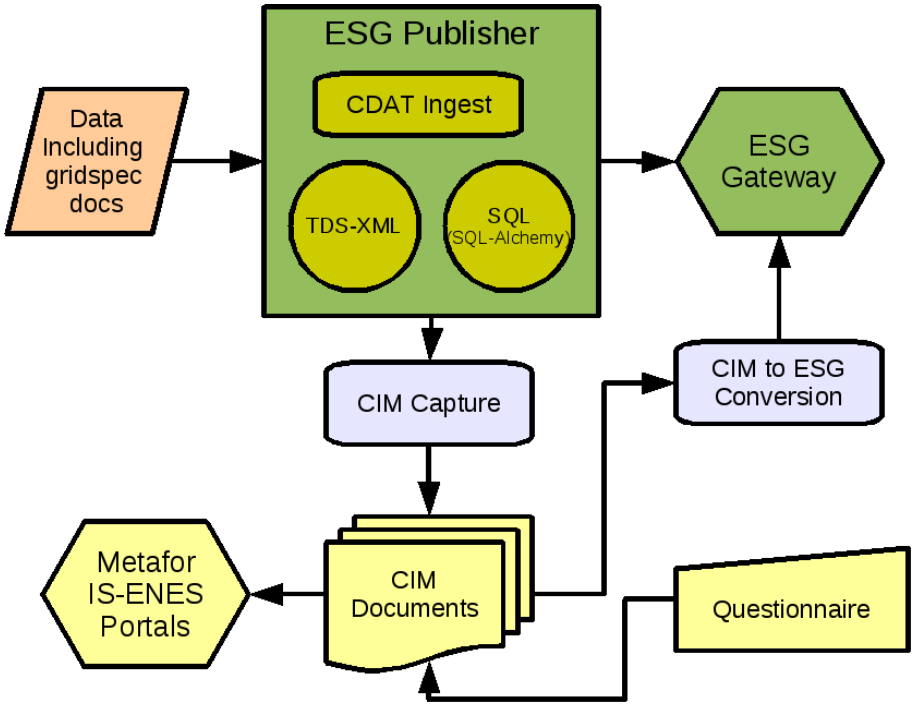

Figure 1: Possible relationship between IS-ENES and ESG portals and their information flow. ESG publisher captures data and raw grid specifications. Metafor captures human input to augment grids, and describe models, simulations and experiments. All information is exposed in both the ESG gateways and future gateways built on CIM content.

An underlying issue for the merge semantics and the relationships between data, metadata, and services is “What is a dataset?”. Some thoughts on that follow from a metafor perspective appear in figure 2.

The key point here is that we need to ensure that the gateway functionality supports providing reusable wget scripts that can get all the data entities within the various datasets of interest, which include at least those in the figure.

Ideally it should be possible for users to share wget scripts, and modify and reuse them themselves.

Action: ESG gateway team to advise on what concept of “dataset” exists in the gateway.

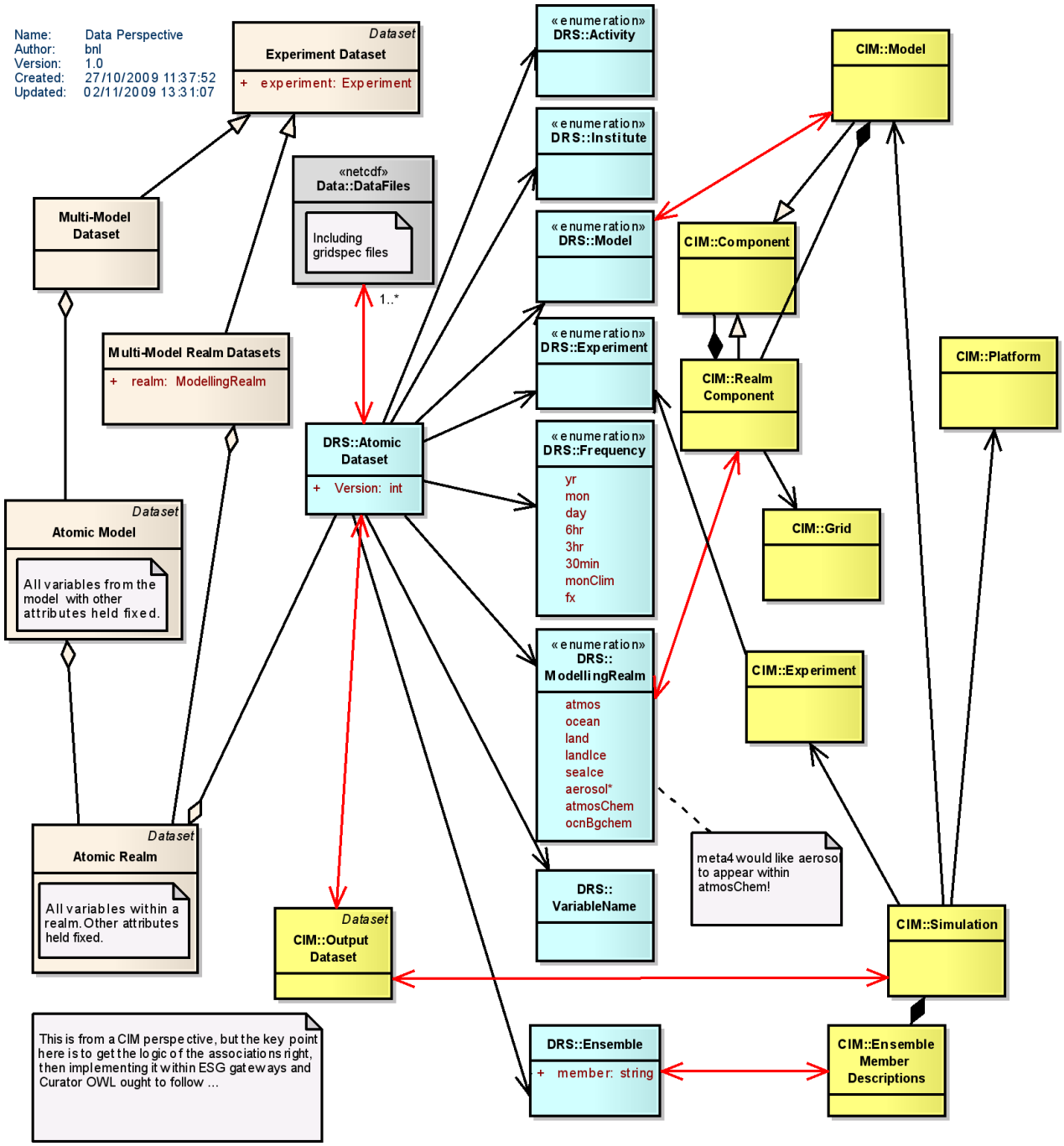

Figure 2: Relationship between key information entities from a DRS and metafor perspective, with a view to what "datasets" users will need to easily access.

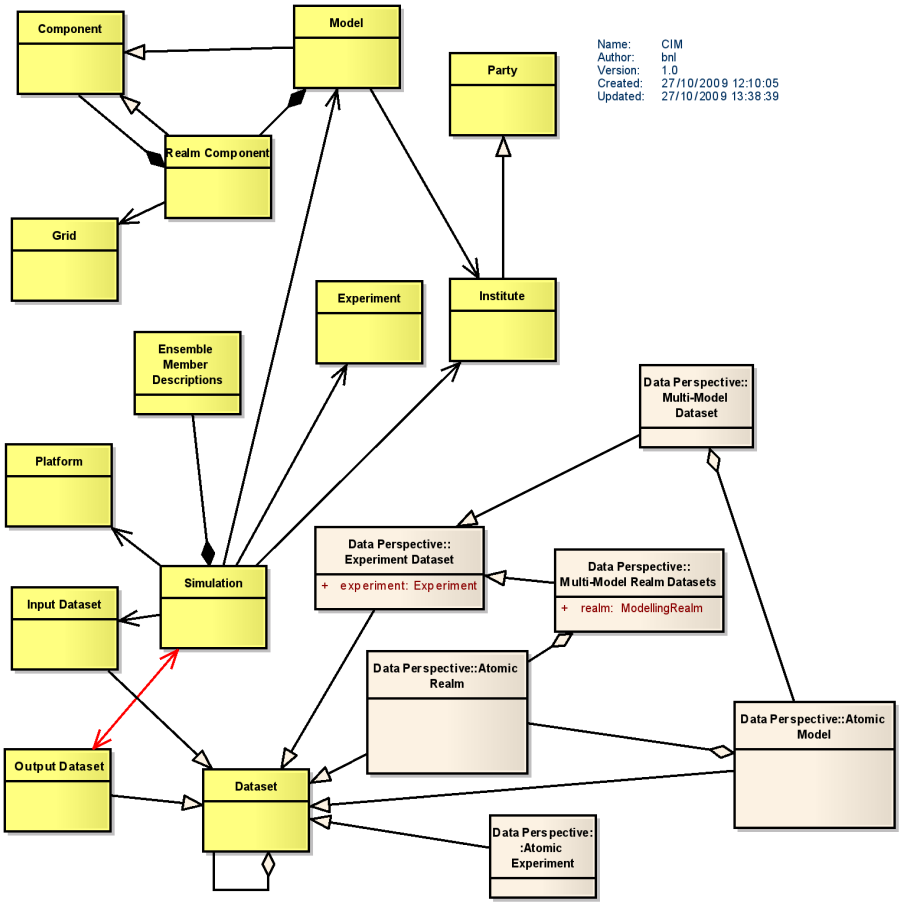

Figure 3: Key CIM entities and their relationships to key dataset concepts.

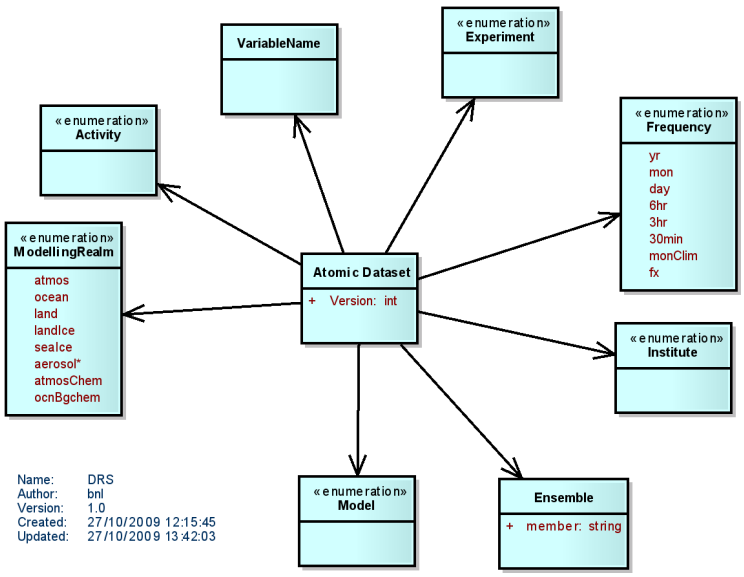

Figure 4: Key DRS entities (as of V20). This will be updated when we have a new proposal for processed data entiies).