A ten minute introduction to ES-DOC technology!

(that might take fifteen minutes)

es-doc
Earth System Documentation

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

SEVENTH FRAMEWORK PROGRAMME

IS-ENES2: FW7 project 312979

Bryan Lawrence

NCAS, STFC & The University of Reading

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

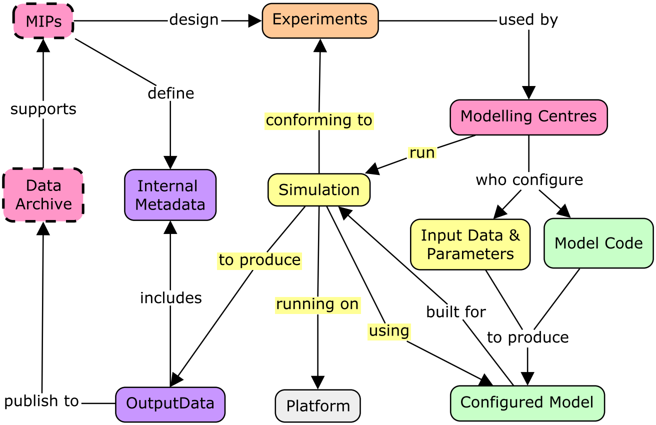
Definitions

- An experiment is an activity aimed at addressing a specific scientific problem.
- We formally describe such an experiment by means of the **NumericalExperiment** which describes the experimental aim, and is composed of a set of **NumericalRequirement**s which need to be met to address the experimental aim, these include any spatio-temporal constraints (what domain is simulated, for how long), forcing constraints (e.g. whether a historical or future scenario is used for anthropogenic emissions of radiatively important gases) etc.

Definitions

- An experiment is an activity aimed at addressing a specific scientific problem.
- We formally describe such an experiment by means of the **NumericalExperiment** which describes the experimental aim, and is composed of a set of **NumericalRequirement**s which need to be met to address the experimental aim, these include any spatio-temporal constraints (what domain is simulated, for how long), forcing constraints (e.g. whether a historical or future scenario is used for anthropogenic emissions of radiatively important gases) etc.
- A **Simulation** is a run of a configured **Model** which conforms to the **NumericalRequirement**s, runs on a **Platform** and produces output **Dataset**s.

Big Picture Workflow



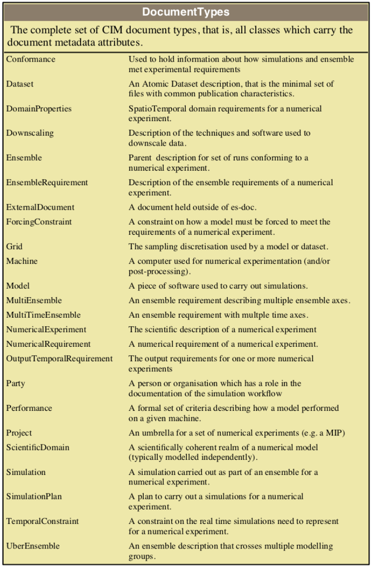
Issues

- Lots of different artefacts created by different individuals at different stages in the workflow.
- Not at all amenable to the traditional “metadata” for “data” paradigm CEDA is used to.
- More in common with the “provenance” work from the computer science community, but
- Much less about automated annotation and more human content generation.

Solution - Documents

es-doc has notion of <<Documents>>, which

- have their own authorship, identity and versioning.
- have their own life-cycle.
- are linked to other documents.
- can be created in many ways, and rendered using many formats. Currently es-doc supports
  - html
  - json
  - python objects (in two different libraries)



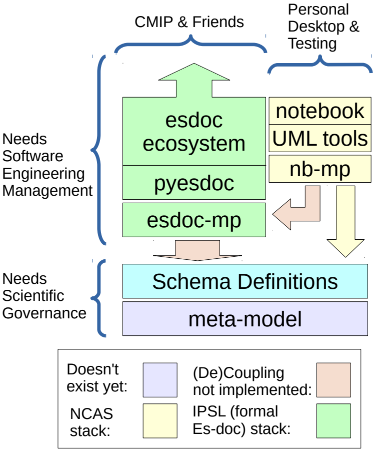
DocumentTypes

The complete set of CIM document types, that is, all classes which carry the document metadata attributes.

Conformance	Used to hold information about how simulations and ensemble met experimental requirements
Dataset	An Atomic Dataset description, that is the minimal set of files with common publication characteristics.
DomainProperties	SpatioTemporal domain requirements for a numerical experiment.
Downscaling	Description of the techniques and software used to downscale data.
Ensemble	Parent description for set of runs conforming to a numerical experiment.
EnsembleRequirement	Description of the ensemble requirements of a numerical experiment.
ExternalDocument	A document held outside of es-doc.
ForcingConstraint	A constraint on how a model must be forced to meet the requirements of a numerical experiment.
Grid	The sampling discretisation used by a model or dataset.
Machine	A computer used for numerical experimentation (and/or post-processing).
Model	A piece of software used to carry out simulations.
MultiEnsemble	An ensemble requirement describing multiple ensemble axes.
MultiTimeEnsemble	An ensemble requirement with multple time axes.
NumericalExperiment	The scientific description of a numerical experiment
NumericalRequirement	A numerical requirement of a numerical experiment.
OutputTemporalRequirement	The output requirements for one or more numerical experiments
Party	A person or organisation which has a role in the documentation of the simulation workflow
Performance	A formal set of criteria describing how a model performed on a given machine.
Project	An umbrella for a set of numerical experiments (e.g. a MIP)
ScientificDomain	A scientifically coherent realm of a numerical model (typically modelled independently).
Simulation	A simulation carried out as part of an ensemble for a numerical experiment.
SimulationPlan	A plan to carry out a simulations for a numerical experiment.
TemporalConstraint	A constraint on the real time simulations need to represent for a numerical experiment.
UberEnsemble	An ensemble description that crosses multiple modelling groups.

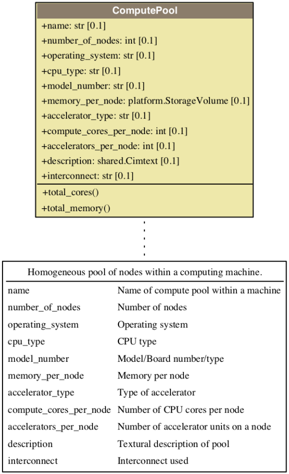
es-doc infrastructure - all python

- All of the basic es-doc concepts are defined using python in a set of schema definitions using a bespoke "esdoc-pythonic-formalism" (which is currently defined in two joint sets of code and a bunch of agreements, it needs a metamode).
- Two independent software stacks exploit those schema (although there is some two-way code which exists but is currently commented-out to avoid dependency hell).


CMIP & Friends
Personal Desktop & Testing
Needs Software Engineering Management
esdoc ecosystem
pyesdoc
esdoc-mp
notebook
UML tools
nb-mp
Needs Scientific Governance
Schema Definitions
meta-model
Doesn't exist yet:
(De)Coupling not implemented:
NCAS stack:
IPSL (formal Es-doc) stack:

Schema Definition Language: ComputePool Example

```
def compute_pool():
    """ Homogeneous pool of nodes within a computing machine. """
    return {
        'type': 'class',
        'base': None,
        'is_abstract': False,
        'properties': [
            ('name', 'str', '0.1',
                'Name of compute pool within a machine'),
            ('number_of_nodes', 'int', '0.1',
                'Number of nodes'),
            ('operating_system', 'str', '0.1',
                'Operating system'),
            ('cpu_type', 'str', '0.1',
                'CPU type'),
            ('model_number', 'str', '0.1',
                'Model/Board number/type'),
            ('memory_per_node', 'platform.storage_volume', '0.1'
                'Memory per node'),
            ('accelerator_type', 'str', '0.1',
                'Type of accelerator'),
            ('compute_cores_per_node', 'int', '0.1',
                'Number of CPU cores per node'),
            ('accelerators_per_node', 'int', '0.1',
                'Number of accelerator units on a node'),
            ('description', 'shared.cimtext', '0.1',
                'Textural description of pool'),
            ('interconnect', 'str', '0.1',
                'Interconnect used'),
            ],
        'derived' : [
            ('total_cores', 'compute_cores_per_node * number_of_
            ('total_memory', 'memory_per_node * number_of_nodes'
```



ComputePool

+name: str [0.1]
+number_of_nodes: int [0.1]
+operating_system: str [0.1]
+cpu_type: str [0.1]
+model_number: str [0.1]
+memory_per_node: platform.StorageVolume [0.1]
+accelerator_type: str [0.1]
+compute_cores_per_node: int [0.1]
+accelerators_per_node: int [0.1]
+description: shared.Cimtext [0.1]
+interconnect: str [0.1]

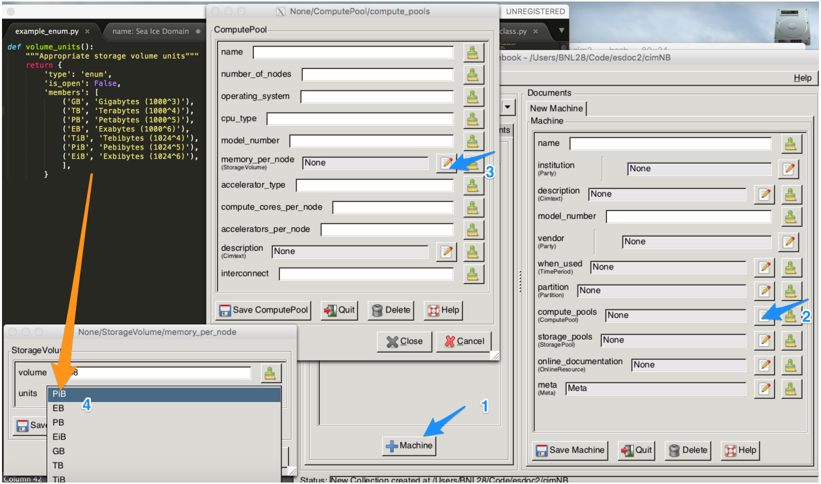
+total_cores()
+total_memory()

Homogeneous pool of nodes within a computing machine.

name	Name of compute pool within a machine
number_of_nodes	Number of nodes
operating_system	Operating system
cpu_type	CPU type
model_number	Model/Board number/type
memory_per_node	Memory per node
accelerator_type	Type of accelerator
compute_cores_per_node	Number of CPU cores per node
accelerators_per_node	Number of accelerator units on a node
description	Textural description of pool
interconnect	Interconnect used

ComputePool

Notebook uses pythonic definitions on the fly



example_enum.py

```
def volume_units():
    """Appropriate storage volume units"""
    return {
        'type': 'enum',
        'is_open': False,
        'members': [
            ('GB', 'Gigabytes (1000^3)'),
            ('TB', 'Terabytes (1000^4)'),
            ('PB', 'Petabytes (1000^5)'),
            ('EB', 'Exabytes (1000^6)'),
            ('TiB', 'Tebibytes (1024^4)'),
            ('PiB', 'Pebibytes (1024^5)'),
            ('EiB', 'Exbibytes (1024^6)'),
            ],
        }
```

None/ComputePool/compute_pools

ComputePool
- name
- number_of_nodes
- operating_system
- cpu_type
- model_number
- memory_per_node (Storage Volume): None
- accelerator_type
- compute_cores_per_node
- accelerators_per_node
- description (Cimtext): None
- interconnect

book - /Users/BNL28/Code/esdoc2/cimNB

Documents / New Machine / Machine
- name
- institution (Party): None
- description (Cimtext): None
- model_number
- vendor (Party): None
- when_used (TimePeriod): None
- partition (Partition): None
- compute_pools (ComputePool): None
- storage_pools (StoragePool): None
- online_documentation (OnlineResource): None
- meta (Meta): Meta

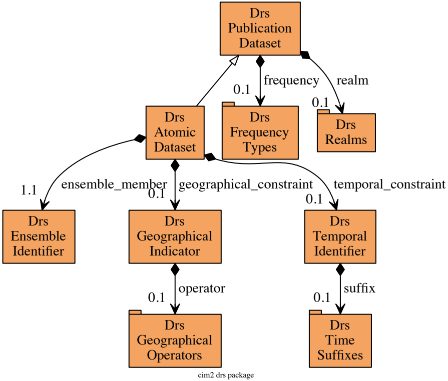
None/StorageVolume/memory_per_node

StorageVolu
- volume: 8
- units: PiB

Status: New Collection created at /Users/BNL28/Code/esdoc2/cimNB

(but the notebook doesn't render the documents yet, waiting on pyesdoc integration for that)

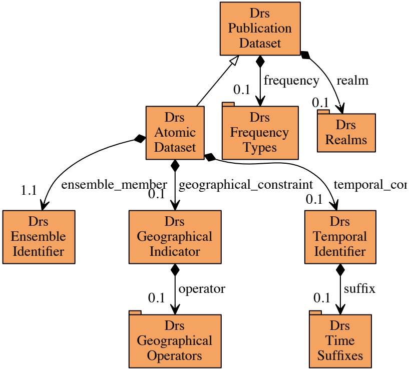
CIM2 packages - DRS example



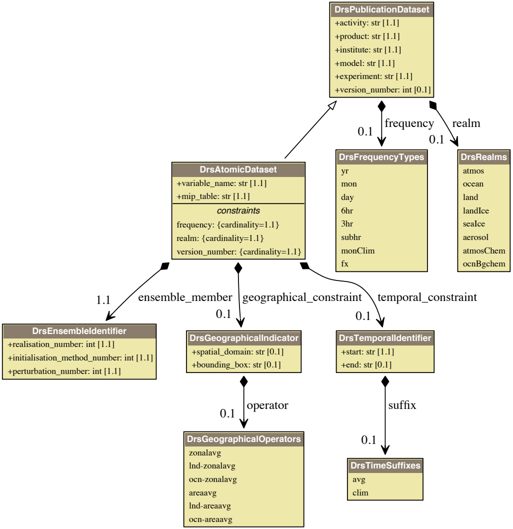
Rectangles = Classes;
Tabs = Enumerations

(The DRS package will change when DRS and file attributes are finalised by the WIP!)

CIM2 packages - DRS example



cim2 drs package

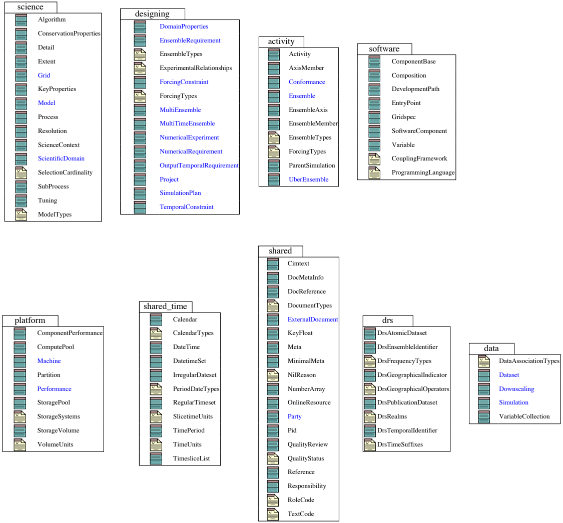


cim2 drs package

Rectangles = Classes;
Tabs = Enumerations

(The DRS package will change when DRS and file attributes are finalised by the WIP!)

CIM2 packages - The complete set



science
Algorithm
ConservationProperties
Detail
Extent
Grid
KeyProperties
Model
Process
Resolution
ScienceContext
ScientificDomain
SelectionCardinality
SubProcess
Tuning
ModelTypes

designing
DomainProperties
EnsembleRequirement
EnsembleTypes
ExperimentalRelationships
ForcingConstraint
ForcingTypes
MultiEnsemble
MultiTimeEnsemble
NumericalExperiment
NumericalRequirement
OutputTemporalRequirement
Project
SimulationPlan
TemporalConstraint

activity
Activity
AxisMember
Conformance
Ensemble
EnsembleAxis
EnsembleMember
EnsembleTypes
ForcingTypes
ParentSimulation
UberEnsemble

software
ComponentBase
Composition
DevelopmentPath
EntryPoint
Gridspec
SoftwareComponent
Variable
CouplingFramework
ProgrammingLanguage

shared
Cimtext
DocMetaInfo
DocReference
DocumentTypes
ExternalDocument
KeyFloat
Meta
MinimalMeta
NilReason
NumberArray
OnlineResource
Party
Pid
QualityReview
QualityStatus
Reference
Responsibility
RoleCode
TextCode

platform
ComponentPerformance
ComputePool
Machine
Partition
Performance
StoragePool
StorageSystems
StorageVolume
VolumeUnits

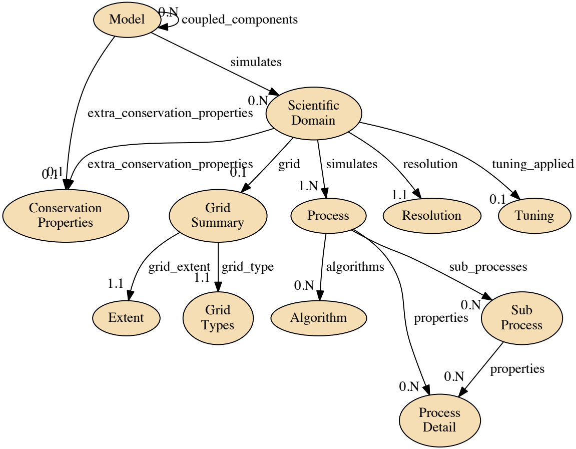
shared_time
Calendar
CalendarTypes
DateTime
DatetimeSet
IrregularDateset
PeriodDateTypes
RegularTimeset
SlicetimeUnits
TimePeriod
TimeUnits
TimesliceList

drs
DrsAtomicDataset
DrsEnsembleIdentifier
DrsFrequencyTypes
DrsGeographicalIndicator
DrsGeographicalOperators
DrsPublicationDataset
DrsRealms
DrsTemporalIdentifier
DrsTimeSuffixes

data
DataAssociationTypes
Dataset
Downscaling
Simulation
VariableCollection

- science
- designing
- activity
- software
- platform
- shared-time, shared
- drs
- data

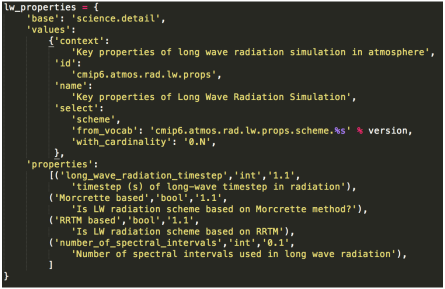
Scientific Descriptions



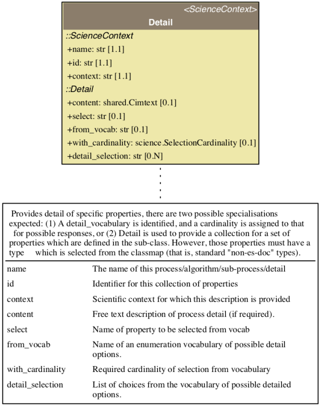
(some minor changes are still underway)

Specialised Extensions

In CMIP5 (CIM1.X) we had “scientific vocabularies” which controlled the properties of some specfic classes.
In CIM2, we have specialisations of the main science classes.



```
lw_properties = {
    'base': 'science.detail',
    'values':
        {'context':
            'Key properties of long wave radiation simulation in atmosphere',
         'id':
            'cmip6.atmos.rad.lw.props',
         'name':
            'Key properties of Long Wave Radiation Simulation',
         'select':
            'scheme',
            'from_vocab': 'cmip6.atmos.rad.lw.props.scheme.%s' % version,
            'with_cardinality': '0.N',
         },
    'properties':
        [('long_wave_radiation_timestep','int','1.1',
             'timestep (s) of long-wave timestep in radiation'),
        ('Morcrette based','bool','1.1',
            'Is LW radiation scheme based on Morcrette method?'),
        ('RRTM based','bool','1.1',
            'Is LW radiation scheme based on RRTM'),
        ('number_of_spectral_intervals','int','0.1',
            'Number of spectral intervals used in long wave radiation'),
        ]
}
```

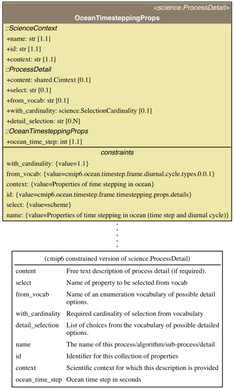


Provides detail of specific properties, there are two possible specialisations expected: (1) A detail_vocabulary is identified, and a cardinality is assigned to that for possible responses, or (2) Detail is used to provide a collection for a set of properties which are defined in the sub-class. However, those properties must have a type which is selected from the classmap (that is, standard "non-es-doc" types).

name	The name of this process/algorithm/sub-process/detail
id	Identifier for this collection of properties
context	Scientific context for which this description is provided
content	Free text description of process detail (if required).
select	Name of property to be selected from vocab
from_vocab	Name of an enumeration vocabulary of possible detail options.
with_cardinality	Required cardinality of selection from vocabulary
detail_selection	List of choices from the vocabulary of possible detailed options.

Detail

Radiation example expanded



<science.ProcessDetail>

OceanTimesteppingProps

::ScienceContext

+name: str [1.1]

+id: str [1.1]

+context: str [1.1]

::ProcessDetail

+content: shared.Cimtext [0.1]

+select: str [0.1]

+from_vocab: str [0.1]

+with_cardinality: science.SelectionCardinality [0.1]

+detail_selection: str [0.N]

::OceanTimesteppingProps

+ocean_time_step: int [1.1]

constraints

with_cardinality: {value=1.1}

from_vocab: {value=cmip6.ocean.timestep.frame.diurnal.cycle.types.0.0.1}

context: {value=Properties of time stepping in ocean}

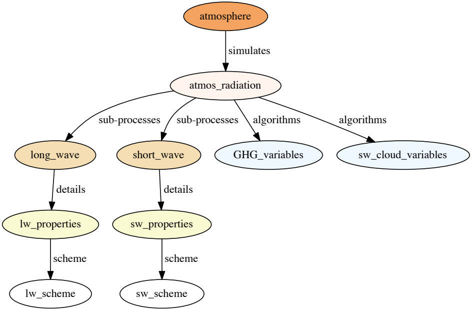
id: {value=cmip6.ocean.timestep.frame.timestepping.props.details}

select: {value=scheme}

name: {value=Properties of time stepping in ocean (time step and diurnal cycle)}

(cmip6 constrained version of science.ProcessDetail)	
content	Free text description of process detail (if required).
select	Name of property to be selected from vocab
from_vocab	Name of an enumeration vocabulary of possible detail options.
with_cardinality	Required cardinality of selection from vocabulary
detail_selection	List of choices from the vocabulary of possible detailed options.
name	The name of this process/algorithm/sub-process/detail
id	Identifier for this collection of properties
context	Scientific context for which this description is provided
ocean_time_step	Ocean time step in seconds

OceanTimesteppingProps



cmip6 (from atmosphere)

(All these figures autogenerated from the definitions.)

Everything else (which is much more)

Sustained effort by Mark Greenslade (IPSL) to ensure that the CIM2 developments will be supportable within the es-doc website and toolchain. Key components will include (but not be limited to):

1. esdoc-py-client: python tools for creating and manipulating documents (and other things)
2. esdoc-shell: command line shell tools for es-doc
3. esdoc-web: software for the esdoc website.
4. esdoc-mp: the "canonical" meta-programming framework
5. esdoc-api: web service API in support of ES-DOC eco-system
6. esdoc-js-client: tool for calling esdoc from javascript

Also major effort by Allyn Treshansky (NOAA):

1. esdoc-questionnaire: tooling for creating documents using a traditional questionnaire technique.

It's worth noting that the Met Office and others will use the esdoc-py-client to directly create CIM2 documents from their workflow metadata database.

References

Notebook and CIM2

- https://bitbucket.org/bnlawrence/esdoc-nb/
- CIM2: In esdoc_nb/mp/core/schema/, moving to it's own package on github next week (I hope).

esdoc toolchain

- Code: https://github.com/ES-DOC/
- Actual working website: https://es-doc.org (CMIP5 metadata mainly)

Lots of activity on slack (ncas-talk.slack.com) in the esdoc channel.

(Health warning: the notebook and scripts currently don't install properly. Some work on python packaging and paths required.)