A ten minute introduction to ES-DOC technology!

(that might take fifteen minutes)

IS-ENES2: FW7 project 312979

Bryan Lawrence

NCAS, STFC & The University of Reading
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 **NumericalRequirements** 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 **NumericalRequirements** 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 **NumericalRequirements**, runs on a **Platform** and produces output **Datasets**.
Big Picture Workflow

- MIPs
  - design
  - supports
  - publish to

- Experiments
  - defined
  - conforming to
  - run
  - who configures

- Modelling Centres

- Data Archive
  - includes

- Internal Metadata
  - to produce
  - running on
  - using

- Simulation

- OutputData
  - to produce

- Platform

- Configured Model
  - to produce

- Input Data & Parameters
  - built for

A ten minute introduction to ES-DOC technology!
Bryan Lawrence - CEDA Vocabulary Meeting, March 1st, 2016
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)
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).
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',
             'Textual description of pool'),
            ('interconnect', 'str', '0.1',
             'Interconnect used'),
        ],
        'derived': [
            ('total_cores', 'compute_cores_per_node * number_of_nodes'),
            ('total_memory', 'memory_per_node * number_of_nodes')
        ]
    }

A ten minute introduction to ES-DOC technology!
Bryan Lawrence - CEDA Vocabulary Meeting, March 1st, 2016
A ten minute introduction to ES-DOC technology!
Bryan Lawrence - CEDA Vocabulary Meeting, March 1st, 2016

Notebook uses pythonic definitions on the fly

(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!)
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
- designing
- activity
- software
- platform
- shared-time, shared
- drs
- data
Scientific Descriptions

(some minor changes are still underway)
In CMIP5 (CIM1.X) we had “scientific vocabularies” which controlled the properties of some specific classes.
In CIM2, we have specialisations of the main science classes.
Radiation example expanded

(All these figures autogenerated from the definitions.)
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.
Notebook and CIM2

- [https://bitbucket.org/bnlawrence/esdoc-nb/](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/](https://github.com/ES-DOC/)
- Actual working website: [https://es-doc.org](https://es-doc.org) (CMIP5 metadata mainly)

Lots of activity on slack ([ncas-talk.slack.com](https://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.)