8th annual
EARTH SYSTEM GRID FEDERATION

PYESSV - A simple controlled vocabulary service

Mark Greenslade (IPSL)
Sebastien Denvil (IPSL)
Guillaume Levavasseur (IPSL)
Atef Ben Nasser (IPSL)

December 5th, 2018

Outline

Introduction
Model
Archive
Web-Service
Usage

Outline

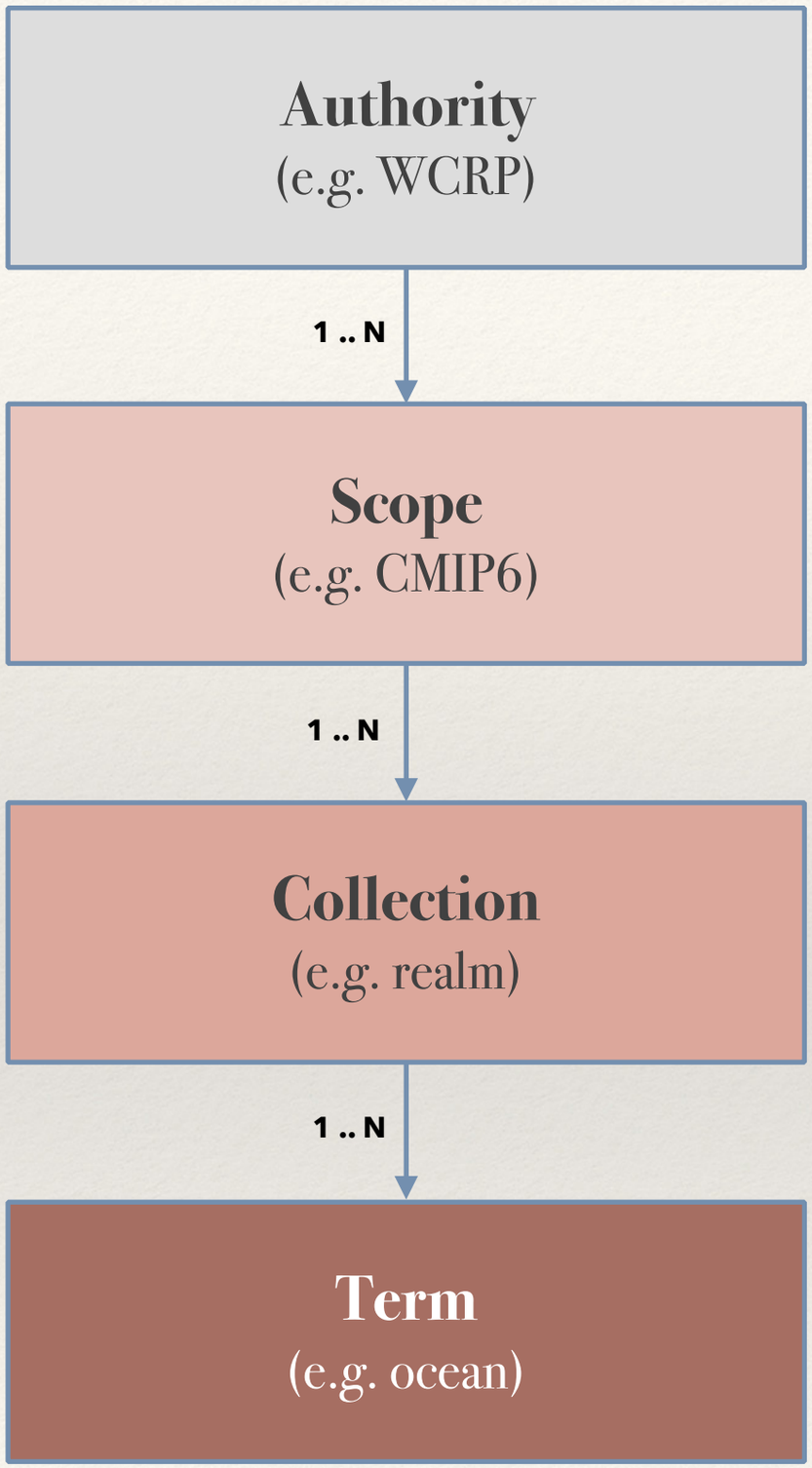
Introduction
Domain Model
Archive
Web-Service
Usage Scenarios

Introduction

Canonical Nameset Micro-Service

- **History**:
  - IPSL Hermes platform 4 yrs ago
- **User Interfaces**:
  - e.g. search facets | visualisations
- **Validation @ Boundaries:**
  - e.g. errata issues
- **Parsing:**
  - e.g. dataset identifiers
- **Driving Scripts:**
  - e.g. cmip6 model spreadsheets
- **Hierarchies:**
  - e.g. institute-id –> source-id

Domain Model



requirements

nodes / edges
hierarchical
associative relationships

design

opinionated
python class model

common attributes

alternative_names
canonical_name
data
description
label
namespace
raw_name
uid
url

Library

https://github.com/es-doc/**pyessv**

- python 2/3
- test driven developed
- OOP & functional hybrid
- github/es-doc/pyessv
- stable - in production

cache
in-memory | file-system

archive
wcrp | esdoc

codecs
dict | json

parsing
strictness levels | templates

i/o
read | write | delete

model
node | iterable-node

factory
type instantiation

Archive

https://github.com/es-doc/**pyessv-archive**

Inputs

WCRP CMIP6 CV's

ESG-F .ini files

ES-DOC scripts

PYESSV WRITERS

Authorities

WCRP

ES-DOC

Scopes

CORDEX

CMIP5

CMIP6

CMIP6

ERRATA

WCRP

CMIP5

CMIP6

CORDEX

activity-id

experiment-id

frequency

institution-id

realm

source-id

source-type

table-id

aerosol

atmos

atmoschem

land

landice

ocean

ocnbgchem

seaice

File System Outputs: 1 MANIFEST file per Authority
1 JSON file per **Term**

```
{
    "_type": "authority",
    "canonical_name": "wcrp",
    "create_date": "2017-06-21 00:00:00+00:00",
    "description": "World Climate Research Program",
    "label": "WCRP",
    "namespace": "wcrp",
    "raw_name": "WCRP",
    "scopes": [
        {
            "_type": "scope",
            "canonical_name": "cmip5",
            "collections": [
                {
                    "_type": "collection",
                    "canonical_name": "cmor-table",
                    "create_date": "2017-06-21 00:00:00+00:00",
                    "description": "ESGF publisher-config CV collection:",
                    "label": "CMOR Table",
                    "namespace": "wcrp:cmip5:cmor-table",
                    "raw_name": "cmor_table",
                    "term_regex": "^[a-z0-9\\-]*$",
                    "terms": [
                        "3hr:3hr",
                        "6hrlev:6hrLev",
                        "6hrplev:6hrPlev",
                        "aero:aero",
                        "amon:Amon",
                        "cf3hr:cf3hr",
                        "cfday:cfDay",
                        "cfmon:cfMon",
                        "cfoff:cfOff",
                        "cfsites:cfSites",
                        "day:day",
                        "fx:fx",
                        "grids:grids",
                        "limon:LImon",
                        "lmon:Lmon",
                        "oclim:Oclim",
                        "oimon:OImon",
                        "omon:Omon",
                        "oyr:Oyr"
                    ],
                    "uid": "e56da3af-9ecc-4fb1-bbaa-0ff477b299f0"
```


~/.esdoc/pyessv-archive/**wcrp**/**MANIFEST**

```
{
    "_type": "term",
    "canonical_name": "amip",
    "create_date": "2017-06-21 00:00:00+00:00",
    "data": {
        "activity_id": [
            "CMIP"
        ],
        "additional_allowed_model_components": [
            "AER",
            "CHEM",
            "BGC"
        ],
        "end_year": "2014",
        "experiment": "AMIP",
        "experiment_id": "amip",
        "min_number_yrs_per_sim": "36",
        "parent_activity_id": [
            "no parent"
        ],
        "parent_experiment_id": [
            "no parent"
        ],
        "required_model_components": [
            "AGCM"
        ],
        "start_year": "1979",
        "sub_experiment_id": [
            "none"
        ],
        "tier": "1"
    },
    "description": "DECK: AMIP",
    "namespace": "wcrp:cmip6:experiment-id:amip",
    "status": "pending",
    "uid": "455bcede-0735-4ebd-9415-6c982ed3274e"
}
```


~/.esdoc/pyessv-archive/**wcrp**/**cmip6**/**experiment-id**/**amip**

Web-Service

https://pyessv.es-doc.org

https://github.com/es-doc/**pyessv-js**

base endpoint:

https://pyessv.es-doc.org/1

retrieval:

GET {base-endpoint}/retrieve/**wcrp**
GET {base-endpoint}/retrieve/**wcrp**/**cmip6**
GET {base-endpoint}/retrieve/**wcrp**/**cmip6**/**experiment-id**
GET {base-endpoint}/retrieve/**wcrp**/**cmip6**/**experiment-id**/**amip**

identifier validation:

GET {base-endpoint}/1/**validate-identifier**
POST {base-endpoint}/1/**validate-identifier-set**

javascript client:

https://github.com/ES-DOC/**pyessv-js**

Usage Scenarios

Usage Scenario: **Dataset Errata**

es-doc
Earth System Documentation

Dataset Errata - Search v0.6.3.0 Support Docs PID Login

Project: CMIP6 | Experiment ID: * | Institution ID: * | Source ID: * | Variable ID: * | Severity: * | Status: *

Total Issues = 20. Filtered Issues = 20.

#	Institute	Title	Created	Updated	Closed	Severity	Status
1	IPSL	200 years extension for piControl	2018-11-29	2018-11-29	--	Low	Resolved
2	IPSL	"Fixed" CMIP6 variables provided by NEMO model are ti ...	2018-11-26	2018-11-27	--	Medium	Resolved
3	NOAA-GFDL	Variable tslsi (3hr,day) has incorrect "comment" vari ...	2018-11-26	--	--	Low	New
4	IPSL	500 years extension for piControl	2018-11-23	2018-11-29	--	Low	Resolved
5	CNRM-CERFACS	Wrong realm ocnBgChem typo	2018-11-14	2018-11-16	--	Low	Resolved
6	NOAA-GFDL	Incorrect some coordinates and cell_methods in piCont ...	2018-11-08	2018-11-08	--	Medium	New
7	NOAA-GFDL	Error in variable "comment" metadata	2018-11-01	2018-11-16	--	Low	New
8	NOAA-GFDL	albisccp erroneous data units	2018-10-29	2018-11-16	--	Low	New
9	IPSL	300 years extension for abrupt-4xCO2	2018-10-22	2018-10-22	--	Low	Resolved
10	IPSL	Irrelevant CFC in experiment other than historical	2018-10-19	2018-10-23	--	Low	Resolved
11	IPSL	Instabilities which lead to erroneous values of tas a ...	2018-10-16	2018-10-16	--	Critical	On Hold
12	IPSL	tas instabilities lead to erroneous values of tasmax	2018-10-05	2018-10-16	--	Critical	On Hold
13	IPSL	Versioning errors for 1pctCO2 and abrupt-4xCO2	2018-07-27	2018-07-27	--	Critical	Resolved
14	IPSL	Wrong realm "ocnBgChm" typo	2018-07-26	2018-08-08	--	Low	Resolved
15	IPSL	Unchanged PIDs for new version	2018-07-20	2018-07-21	--	High	Resolved
16	IPSL	Some sea ice variables in 3D instead of 1D	2018-07-12	2018-07-17	--	Low	Resolved
17	IPSL	Time instantaneous data with time boundaries	2018-07-02	2018-11-29	--	Low	Wont Fix
18	IPSL	Integers instead of PFTs names	2018-07-02	2018-10-12	--	Low	Resolved
19	IPSL	Integers instead of ocean passages names	2018-07-02	2018-07-17	--	Low	Resolved
20	IPSL	"area:coordinates" attribute is missing	2018-07-02	2018-07-17	--	Low	Resolved

Total Issues = 20. Filtered Issues = 20. v0.6.3.0 © ES-DOC

vocabularies

pid-task-action
pid-task-status
project
issue-severity
issue-status

front-end

leverages pyessv-js

server-side

dataset identifier validation
errata field validation

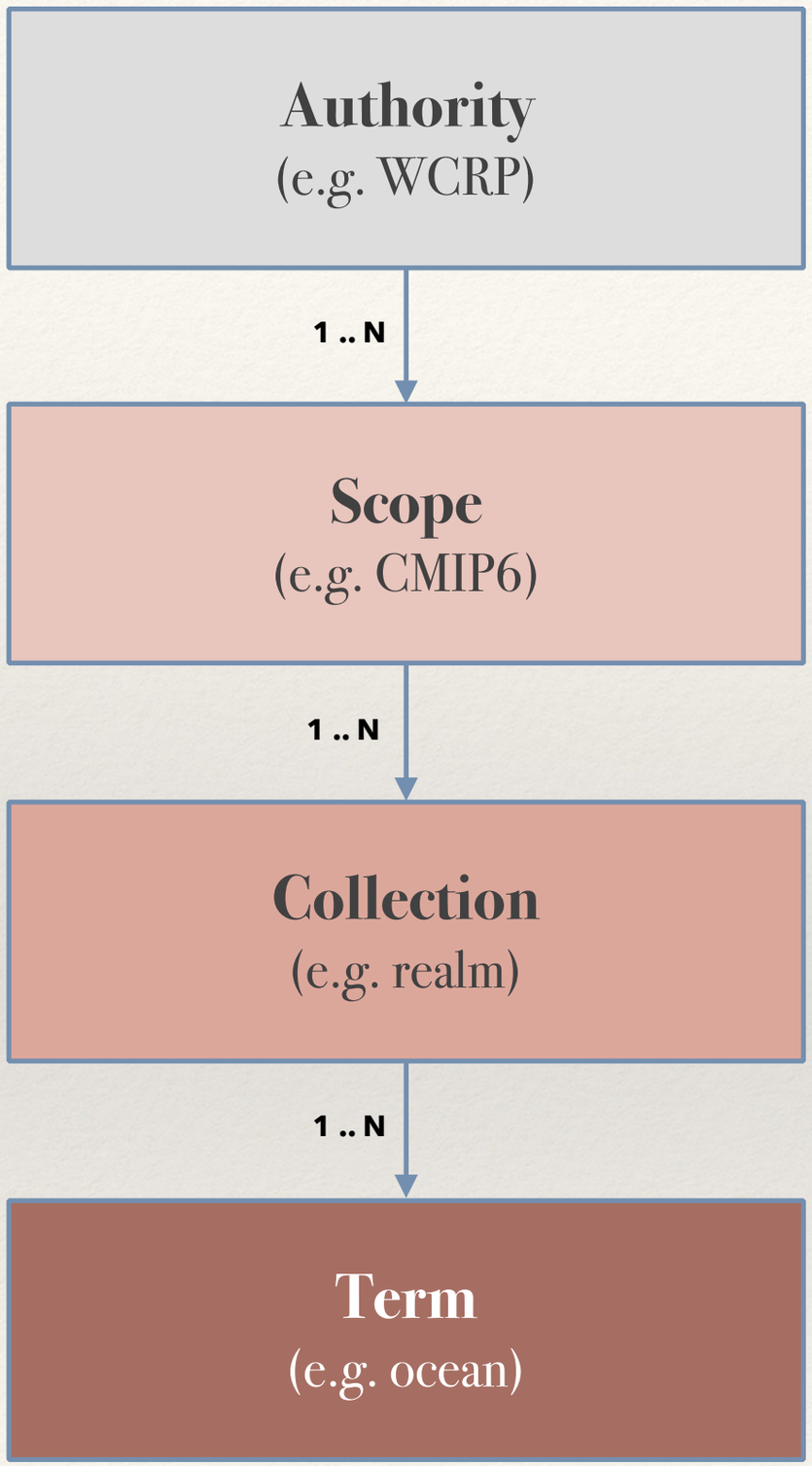
Usage Scenario: **Dataset Errata**

```
from pyessv._factory import create_template_parser
from pyessv._constants import PARSING_STRICTNESS_1

# Template that identifiers must conform to.
_TEMPLATE = 'CMIP6.{}.{}.{}.{}.{}.{}.{}.{}'

# Collections injected into template.
_COLLECTIONS = (
    'wcrp:cmip6:activity-id',
    'wcrp:cmip6:institution-id',
    'wcrp:cmip6:source-id',
    'wcrp:cmip6:experiment-id',
    'wcrp:cmip6:member-id',
    'wcrp:cmip6:table-id',
    'wcrp:cmip6:variable-id',
    'wcrp:cmip6:grid-label'
    )

def parse(identifier):
    """Parses a CMIP6 dataset identifier.

    """
    parser = create_template_parser(_TEMPLATE, _COLLECTIONS, PARSING_STRICTNESS_1)

    # Strip version suffix.
    if '#' in identifier:
      identifier = identifier.split('#')[0]

    return parser.parse(identifier)
```

Usage Scenario: **Documenting Models**

```
# Import pyessv - auto-loads archive.
import pyessv

import init_from_cmip5
import write_pdf
import write_xls


# Processing pipeline.
PIPELINE = (init_from_cmip5, write_xls, write_pdf)


"""Yields vocabulary vectors to push through pipeline."""
def yield_vectors():
    for i in pyessv.WCRP.cmip6.institution_id:
        for s in pyessv.WCRP.cmip6.get_institute_sources(i):
            for t in pyessv.ESDOC.cmip6.get_model_topics(s):
                yield i, s, t


"""Execute model documentation pipeline."""
def main():
    for i, s, r in yield_vectors():
        for task in PIPELINE:
            task(i, s, r)
```

Summary



present

in production
pragmatic
simple
extensible
decent set of features

future

SKOS codec
documentation
new scopes
new contributors
CI server

8th annual
EARTH SYSTEM GRID FEDERATION

PYESSV - A simple controlled vocabulary service

Mark Greenslade (IPSL)
Sebastien Denvil (IPSL)
Guillaume Levavasseur (IPSL)
Atef Ben Nasser (IPSL)

December 5th, 2018

Outline

Introduction
Model
Archive
Web-Service
Usage