PYESSV - A simple controlled vocabulary service

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

Outline
Introduction
Model
Archive
Web-Service
Usage

December 5th, 2018
Outline

Introduction
Domain Model
Archive
Web-Service
Usage Scenarios
Introduction
Canonical Nameset Micro-Service
Introduction

<table>
<thead>
<tr>
<th>Model</th>
<th>Library</th>
<th>Archive</th>
<th>Web-Service</th>
<th>Usage</th>
</tr>
</thead>
</table>

‣ 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
Authority (e.g. WCRP)

Scope (e.g. CMIP6)

Collection (e.g. realm)

Term (e.g. ocean)

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
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-zA-Z0-9\-\_]+$",  
          "terms": [  
            "3hr:3hr",  
            "6hrlev:6hrLev",  
            "6hrplev:6hrPlev",  
            "aero:aero",  
            "amon:Ammon",  
            "cf3hr:cf3hr",  
            "cfday:cfDay",  
            "cfmon:cfMon",  
            "cfoff:cfOff",  
            "cfsites:cfSites",  
            "day:day",  
            "fx:fx",  
            "grids:grids",  
            "limon:Limon",  
            "limon:Limon",  
            "oclim:oclim",  
            "omin:Omon",  
            "omon:Omon",  
            "dyr:Yr"  
          ]  
        }  
      ]  
    }  
  ]  
  "uid": "e56da3af-9ecc-4f81-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": "455bcde-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

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

**front-end**
- leverages pyessv-js

**server-side**
- dataset identifier validation
- errata field validation

<table>
<thead>
<tr>
<th>#</th>
<th>Institute</th>
<th>Title</th>
<th>Created</th>
<th>Updated</th>
<th>Closed</th>
<th>Severity</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PSL</td>
<td>200 years extension for pControl</td>
<td>2018-11-29</td>
<td>2018-11-29</td>
<td>--</td>
<td>Low</td>
<td>Resolved</td>
</tr>
<tr>
<td>2</td>
<td>PSL</td>
<td>“Fixed” CMIP variables provided by NEMO model are...</td>
<td>2018-11-26</td>
<td>2018-11-27</td>
<td>--</td>
<td>Medium</td>
<td>Resolved</td>
</tr>
<tr>
<td>3</td>
<td>NOAA-GFDL</td>
<td>Variable test (Brudding) has incorrect “comment” vari...</td>
<td>2018-11-26</td>
<td>--</td>
<td>--</td>
<td>Low</td>
<td>New</td>
</tr>
<tr>
<td>4</td>
<td>PSL</td>
<td>500 years extension for pControl</td>
<td>2018-11-23</td>
<td>2018-11-29</td>
<td>--</td>
<td>Low</td>
<td>Resolved</td>
</tr>
<tr>
<td>5</td>
<td>CNRM-CERFACS</td>
<td>Wrong realm con/BioChem typo</td>
<td>2018-11-14</td>
<td>2018-11-16</td>
<td>--</td>
<td>Low</td>
<td>Resolved</td>
</tr>
<tr>
<td>6</td>
<td>NOAA-GFDL</td>
<td>Incorrect some coordinates and cell_methods in pCont...</td>
<td>2018-11-08</td>
<td>2018-11-08</td>
<td>--</td>
<td>Medium</td>
<td>New</td>
</tr>
<tr>
<td>7</td>
<td>NOAA-GFDL</td>
<td>Error in variable “comment” metadata</td>
<td>2018-11-01</td>
<td>2018-11-16</td>
<td>--</td>
<td>Low</td>
<td>New</td>
</tr>
<tr>
<td>8</td>
<td>NOAA-GFDL</td>
<td>abisvol erroneous data units</td>
<td>2018-10-29</td>
<td>2018-11-16</td>
<td>--</td>
<td>Low</td>
<td>New</td>
</tr>
<tr>
<td>9</td>
<td>PSL</td>
<td>300 years extension for abrupt-4xCO2</td>
<td>2018-10-22</td>
<td>2018-10-22</td>
<td>--</td>
<td>Low</td>
<td>Resolved</td>
</tr>
<tr>
<td>10</td>
<td>PSL</td>
<td>Irrelevant CFC in experiment other than historical</td>
<td>2018-10-19</td>
<td>2018-10-23</td>
<td>--</td>
<td>Low</td>
<td>Resolved</td>
</tr>
<tr>
<td>11</td>
<td>PSL</td>
<td>Instabilities which lead to erroneous values of tas a...</td>
<td>2018-10-16</td>
<td>2018-10-16</td>
<td>--</td>
<td>Critical</td>
<td>On Hold</td>
</tr>
<tr>
<td>12</td>
<td>PSL</td>
<td>tas instabilities lead to erroneous values of temmax</td>
<td>2018-10-05</td>
<td>2018-10-16</td>
<td>--</td>
<td>Critical</td>
<td>On Hold</td>
</tr>
<tr>
<td>13</td>
<td>PSL</td>
<td>Versioning errors for 1pc002 and abrupt-4xCO2</td>
<td>2018-07-27</td>
<td>2018-07-27</td>
<td>--</td>
<td>Critical</td>
<td>Resolved</td>
</tr>
<tr>
<td>14</td>
<td>PSL</td>
<td>Wrong realm “con/BioChem” typo</td>
<td>2018-07-26</td>
<td>2018-08-08</td>
<td>--</td>
<td>Low</td>
<td>Resolved</td>
</tr>
<tr>
<td>15</td>
<td>PSL</td>
<td>Unchanged PIDs for new version</td>
<td>2018-07-20</td>
<td>2018-07-21</td>
<td>--</td>
<td>High</td>
<td>Resolved</td>
</tr>
<tr>
<td>16</td>
<td>PSL</td>
<td>Some sea ice variables in 3D instead of 1D</td>
<td>2018-07-12</td>
<td>2018-07-17</td>
<td>--</td>
<td>Low</td>
<td>Resolved</td>
</tr>
<tr>
<td>17</td>
<td>PSL</td>
<td>Time instantaneous data with time boundaries</td>
<td>2018-07-02</td>
<td>2018-11-29</td>
<td>--</td>
<td>Low</td>
<td>Work Fix</td>
</tr>
<tr>
<td>18</td>
<td>PSL</td>
<td>Integers instead of PFTs names</td>
<td>2018-07-02</td>
<td>2018-10-12</td>
<td>--</td>
<td>Low</td>
<td>Resolved</td>
</tr>
<tr>
<td>19</td>
<td>PSL</td>
<td>Integers instead of ocean passages names</td>
<td>2018-07-02</td>
<td>2018-07-17</td>
<td>--</td>
<td>Low</td>
<td>Resolved</td>
</tr>
<tr>
<td>20</td>
<td>PSL</td>
<td>“Area coordinates” attribute is missing</td>
<td>2018-07-02</td>
<td>2018-07-17</td>
<td>--</td>
<td>Low</td>
<td>Resolved</td>
</tr>
</tbody>
</table>

Total Issues = 20, Filtered Issues = 20.
Usage Scenario: **Dataset Errata**

```python
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**

```python
# 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
**Authority** (e.g. WCRP)

**Scope** (e.g. CMIP6)

**Collection** (e.g. realm)

**Term** (e.g. ocean)

---

**present**
- in production
- pragmatic
- simple
- extensible
- decent set of features

**future**
- SKOS codec documentation
- new scopes
- new contributors
- CI server
PYESSV - A simple controlled vocabulary service

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

Outline
Introduction
Model
Archive
Web-Service
Usage