



Koninklijk Nederlands  
Meteorologisch Instituut  
*Ministerie van Infrastructuur en Milieu*

# **SWIRRL**

## **Provenance-Aware Reproducible Workspaces**

**Alessandro Spinuso**, Ian van der Neut  
Hans Verhoef, Friedrich Striewski, Mats Veldhuizen

R&D Data Technology and Observations



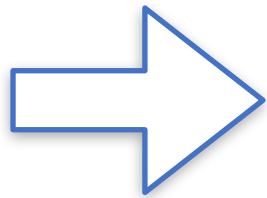
## Objective: Extend C4I with Data Driven & Reproducible Workspaces

Climate4Impact Search for CMIP5/6  
Cordex Data (Distributed Data)

Incremental data staging/subsetting onto customisable  
and Reproducible Notebooks (extensible to other tools..)

The screenshot shows a search interface with filters for frequency (day), experiment (ssp585), and source (EC-Earth3). It displays several categories of climate parameters:

- Temperature:** ta - Air temperature (104), tas - Temperature (90), tasmin - Min. Temperature (81), tasmax - Max. Temperature (81)
- Precipitation:** pr - Precipitation (90), prsn - Snow (72), prc - Convective precipitation (71)
- Humidity:** hurs - Rel. Humidity (79), huss - Specific humidity (74), rhsm - Min. Rel. Humidity (-), rhs - Rel. Humidity (-), hus - Spec. Humidity (54), hur - Rel. Humidity (22)
- Wind:** uas - Eastward wind (74), vas - Northward wind (74), sfcWind - Wind (72), sfcWindmax - Max Wind (31)
- Radiation:** rads - SW Radiation Dn (72), rlds - LW Radiation Dn (72), rsus - SW Radiation Up (22), rfls - LW Radiation Up (22), radsdiff - Diff. Radiation (-), clt - Cloud (22)
- Pressure:** ps - Pressure (-), psl - Sea level pressure (79), pfull - Pressure (-)
- Evaporation:** evspsbl - Act. Evap. (-), evsblpot - Pot. Evap. (-), evspsblsol - Sol Evap. (-), evspsblveg - Canopy Evap. (-)



The screenshot shows search results for project CMIP6, variable ta, and variable tas. A modal dialog is open, indicating that 100 datasets have been selected and are ready for download. The dialog text reads: "File list ready for download (limited to 100 datasets). Found 100 datasets containing 518 file links." There are buttons for "GET LIST AS JSON" and "CLOSE".

# Interactive & reproducible Workspaces



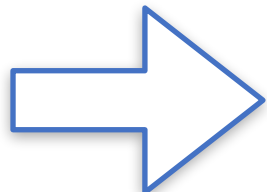
**Objective: Extend C4I with Data Driven & Reproducible Workspaces**

Climate4Impact Search for CMIP5/6  
Cordex Data (Distributed Data)

Incremental data staging/subsetting onto customisable  
and Reproducible Notebooks (extensible to other tools..)

The screenshot shows a search interface with filters for frequency (day), experiment (ssp585), and source (EC-Earth3). It displays several category cards with checkboxes for parameters:

- Temperature:** ta - Air temperature (104), tas - Temperature (90), tasmin - Min. Temperature (81), tasmax - Max. Temperature (81)
- Precipitation:** pr - Precipitation (90), prsn - Snow (72), prc - Convective precipitation (71)
- Humidity:** hurs - Rel. Humidity (79), huss - Specific humidity (74), rhsm - Min. Rel. Humidity (-), rhs - Rel. Humidity (-), hus - Spec. Humidity (84), hur - Rel. Humidity (22)
- Wind:** uas - Eastward wind (74), vas - Northward wind (74), sfcWind - Wind (72), sfcWindmax - Max Wind (31)
- Radiation:** rads - SW Radiation Dn (72), rlds - LW Radiation Dn (72), rsus - SW Radiation Up (22), rfls - LW Radiation Up (22), radsdiff - Diff. Radiation (-), clt - Cloud (22)
- Pressure:** ps - Pressure (-), psl - Sea level pressure (79), pfull - Pressure (-)
- Evaporation:** evspsbl - Act. Evap. (-), evsblpot - Pot. Evap. (-), evspsblsol - Sol Evap. (-), evspsblveg - Canopy Evap. (-)



The screenshot shows a Jupyter Notebook environment with a file browser on the left displaying search results for 'tasmax\_day\_EC-Earth3\_ssp585\_r11p1f1\_gr\_20230101-20231231.nc'. The notebook cell contains the following code:

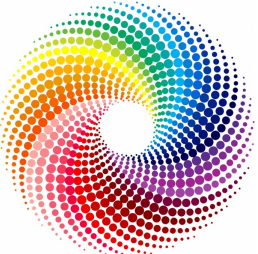
```

[1]: from icclim import icclim
import numpy as np
import netCDF4
import matplotlib.pyplot as plt
import matplotlib
import sys
import glob
import os
import datetime
import cftime

print("python: ", sys.version)
print("numpy: ", np.__version__)
print("netCDF4: ", netCDF4.__version__)
print("matplotlib: ", matplotlib.__version__)

python: 3.6.11 | packaged by conda-forge | (default, Aug 5 2020, 20:09:42)
[GCC 7.5.0]

```



SWIRRL-API

- Trace Changes to Software and Data
- Restore Environments

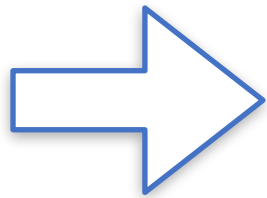
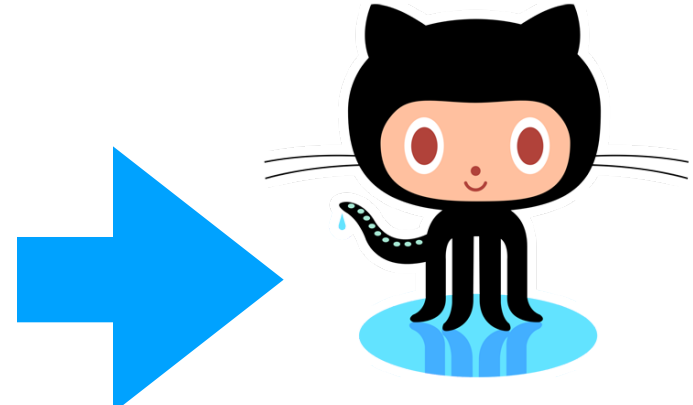
# Interactive & reproducible Workspaces



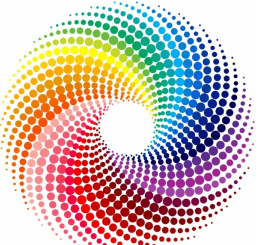
**Objective: Extend C4I with Data Driven & Reproducible Workspaces**

Climate4Impact Search for CMIP5/6  
Cordex Data (Distributed Data)

Incremental data staging/subsetting onto customisable  
and Reproducible Notebooks (extensible to other tools..)

Software and Environment to Git



SWIRRL-API

- Trace Changes to Software and Data
- Restore Environments

# Interactive & reproducible Workspaces



**Objective: Extend C4I with Data Driven & Reproducible Workspaces**

Climate4Impact Search for CMIP5/6  
Cordex Data (Distributed Data)

Incremental data staging/subsetting onto customisable  
and Reproducible Notebooks (extensible to other tools..)

The screenshot shows a search interface with filters for PARAMETER, FREQUENCY, EXPERIMENT, and MODEL. It displays several categories of climate variables with checkboxes and counts:

- Temperature:** ta - Air temperature (104), tas - Temperature (90), tasmin - Min. Temperature (81), tasmax - Max. Temperature (81)
- Precipitation:** pr - Precipitation (90), prsn - Snow (72), prc - Convective precipitation (71)
- Humidity:** hurs - Rel. Humidity (79), huss - Specific humidity (74), rhsm - Min. Rel. Humidity (-), rhs - Rel. Humidity (-), hus - Spec. Humidity (84), hur - Rel. Humidity (22)
- Wind:** uas - Eastward wind (74), vas - Northward wind (74), sfcWind - Wind (72), sfcWindmax - Max Wind (31)
- Radiation:** rds - SW Radiation Dn (72), rlds - LW Radiation Dn (72), rsus - SW Radiation Up (22), rfls - LW Radiation Up (22), rdsdiff - Diff. Radiation (-), clt - Cloud (22)
- Pressure:** ps - Pressure (-), psl - Sea level pressure (79), pfull - Pressure (-)
- Evaporation:** evspsbl - Act. Evap. (-), evsblpot - Pot. Evap. (-), evspsblsoi - Sol Evap. (-), evspsblveg - Canopy Evap. (-)

The screenshot shows a Jupyter Notebook environment. On the left, a 'Search results' window lists files from a project named 'CMIP6', including files like 'swirrl\_fileinfo.json' and various 'tasmax\_day\_EC-Earth3\_ssp585\_r11p1f1\_gr\_...' files. The main notebook area shows a code cell titled 'ICCLIM C4I Demo' with the following code:

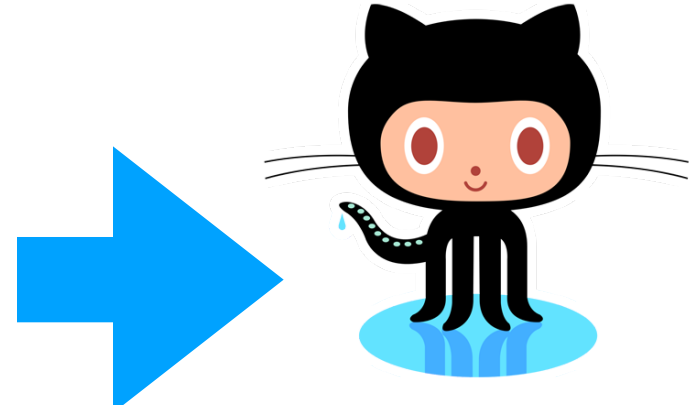
```

[1]: from icclim import icclim
import numpy as np
import netCDF4
import matplotlib.pyplot as plt
import matplotlib
import sys
import glob
import os
import datetime
import cftime

print("python: ", sys.version)
print("numpy: ", np.__version__)
print("netCDF4: ", netCDF4.__version__)
print("matplotlib: ", matplotlib.__version__)

python: 3.6.11 | packaged by conda-forge | default, Aug 5 2020, 20:09:42
[GCC 7.5.0]

```



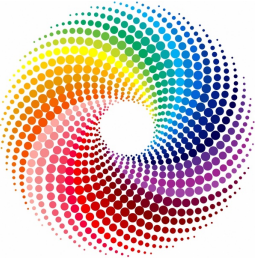
Software and Environment to Git



MyBinder Reproduce

- Trace Changes to Software and Data
- Restore Environments

Data



SWIRRL-API

# Provenance-aware Workspaces

## SWIRRL-API



**A Web API (high-level piece of infrastructure) to:**

Manage **Working Sessions** offering **Notebook and Visualisation Services**

Run Workflows (CWL) for data staging and preprocessing onto the Working Session

Keep data staging history

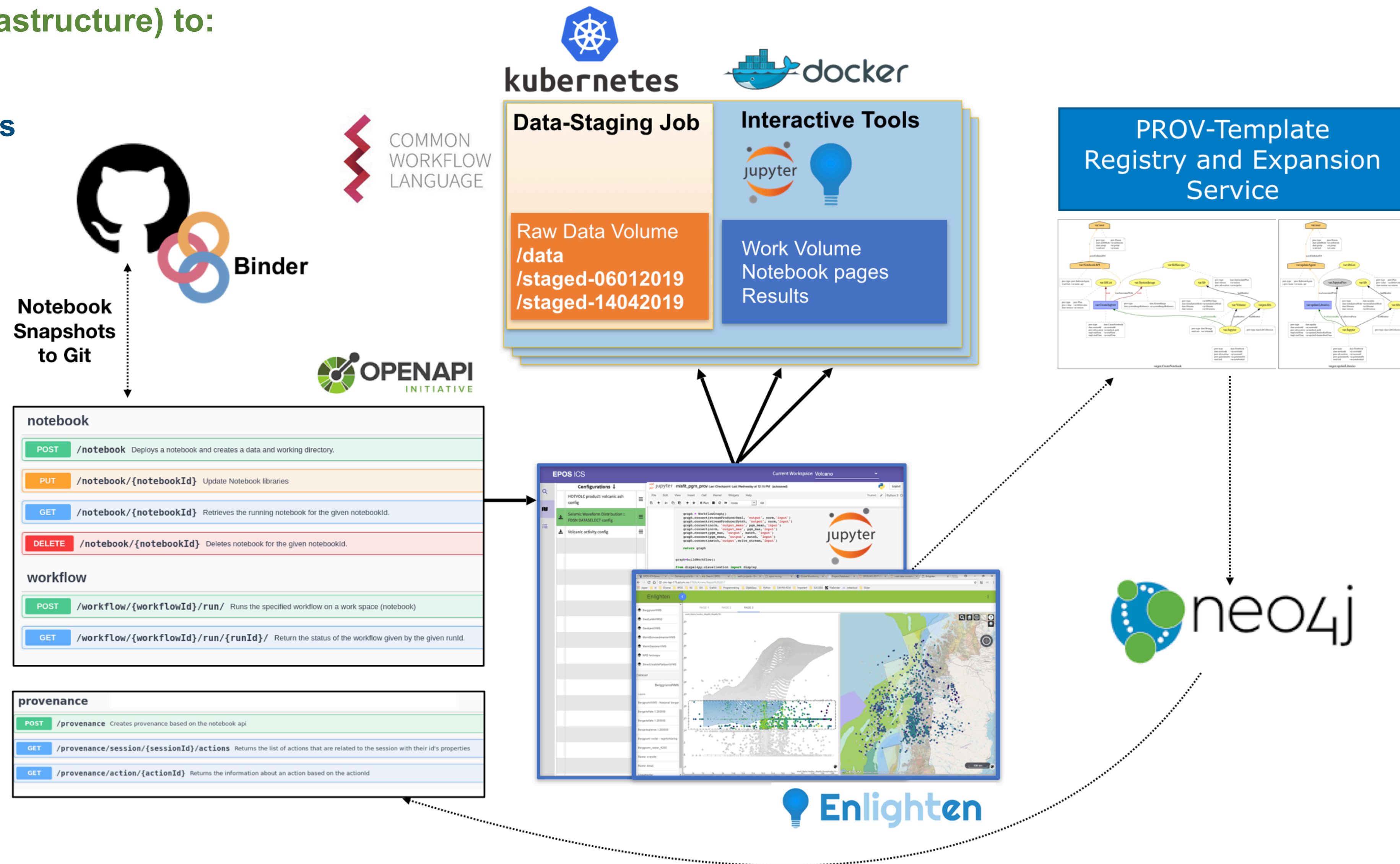
Provenance-aware

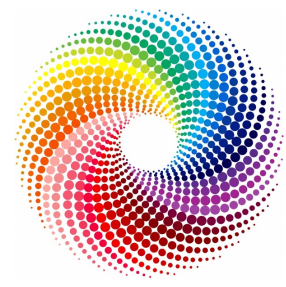
Restore SW Environments to to a state in the past

Ondemand Binder Snapshots to GitHub (Environment, methods, data references)

**Jupyter Lab Extension!**

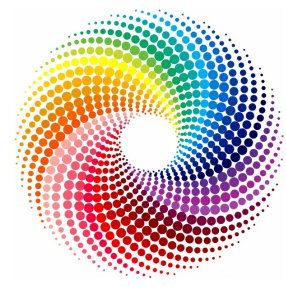
<https://gitlab.com/KNMI-OSS/swirrl/swirrl-api>  
[https://zenodo.org/record/4264852#.X7ZeqNv\\_qNZ](https://zenodo.org/record/4264852#.X7ZeqNv_qNZ)



**notebook**

provide/manage a customisable notebook environment

**POST** **/notebook** Deploys a notebook and creates a data and working directory.**POST** **/notebook/{id}/snapshot** Creates a notebook snapshot**PUT** **/notebook/{id}/restorelibs/{activityId}** Restore Notebook libraries from a previous update**PUT** **/notebook/{id}** Update Notebook libraries**GET** **/notebook/{id}** Retrieves the running notebook for the given notebookId.**DELETE** **/notebook/{id}** Deletes notebook for the given notebookId.



## notebook

provide/manage a customisable notebook environment



**POST** /notebook Deploys a notebook and creates a data and working directory.

**POST** /notebook/{id}/snapshot Creates a notebook snapshot

**PUT** /notebook/{id}/restorelibs/{activityId} Restore Notebook libraries from a previous update

**PUT** /notebook/{id} Update Notebook libraries

**GET** /notebook/{id} Retrieves the running notebook for the given notebookId.

**DELETE** /notebook/{id} Deletes notebook for the given notebookId.

## workflow

data staging / data preparation / processing



**POST** /workflow/{workflowId}/run/ Runs the specified workflow on a work space (notebook)

**GET** /workflow/{workflowId}/run/{runId}/ Return the status of the workflow given by the given runId.



COMMON  
WORKFLOW  
LANGUAGE



docker





# “Update” a Notebook Environment

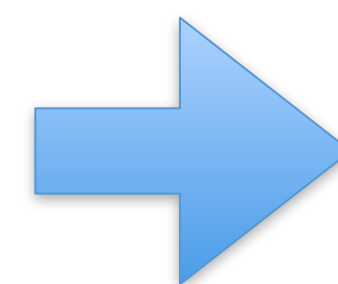


```
document
  prefix vargen <http://openprovenance.org/vargen#>
  prefix s-prov <http://s-prov/ns/#>
  prefix pre_0 <http://www.w3.org/2001/XMLSchema>
  prefix dare <http://project-dare.eu/ns#>
  prefix d-prov <http://d-prov.org/#>
  prefix dcterms <http://purl.org/dc/terms/>
  prefix vcard <http://www.w3.org/2006/vcard/ns#>
  prefix var <http://openprovenance.org/var#>
  prefix tpl <http://openprovenance.org/tmpl#>
  prefix foaf <http://xmlns.com/foaf/0.1/>
  prefix uuid <urn:uuid:>

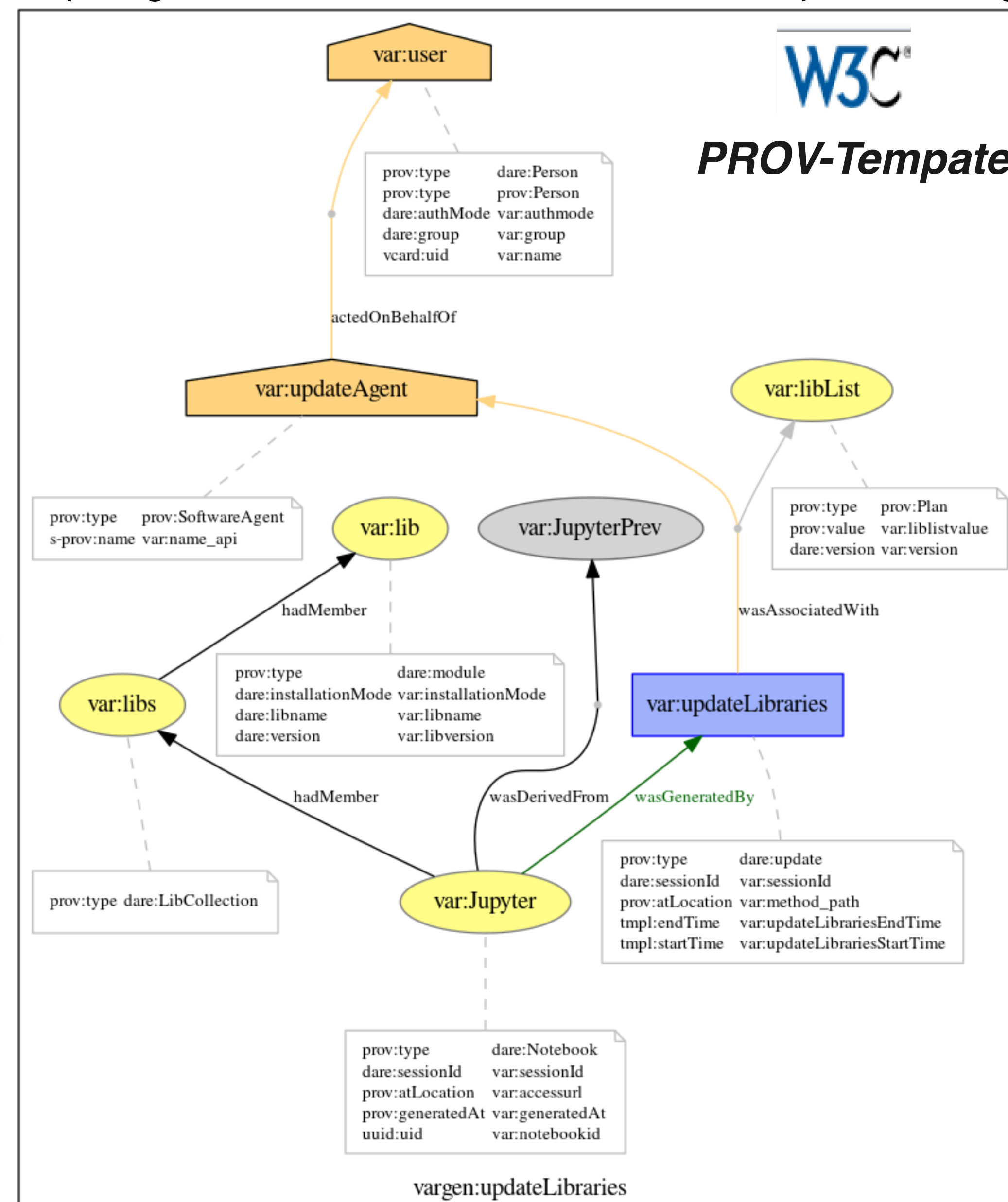
  bundle vargen:updateLibraries
    prefix vargen <http://openprovenance.org/vargen#>
    prefix s-prov <http://s-prov/ns/#>
    prefix dare <http://project-dare.eu/ns#>
    prefix vcard <http://www.w3.org/2006/vcard/ns#>
    prefix var <http://openprovenance.org/var#>
    prefix tpl <http://openprovenance.org/tmpl#>
    prefix dcterms <http://purl.org/dc/terms/>
    prefix uuid <urn:uuid:>

    entity(var:Jupyter, [prov:generatedAt='var:generatedAt', uuid:uid='var:notebookid', prov:atLocat
    entity(var:lib, [dare:libname='var:libname', dare:installationMode='var:installationMode', dare:
    entity(var:libList, [dare:version='var:version', prov:type='prov:Plan', prov:value='var:liblistv
    entity(var:libs, [prov:type='dare:LibCollection'])
    wasDerivedFrom(var:Jupyter, var:JupyterPrev, -, -, -)
    wasAssociatedWith(var:updateLibraries, var:updateAgent, var:libList)
    activity(var:updateLibraries, -, -, [prov:atLocation='var:method_path', tpl:startTime='var:upda
    actedOnBehalfOf(var:updateAgent, var:user, -)
    wasGeneratedBy(var:Jupyter, var:updateLibraries, -)
    agent(var:user, [vcard:uid='var:name', dare:authMode='var:authmode', dare:group='var:group', pro
    agent(var:updateAgent, [prov:type='prov:SoftwareAgent', s-prov:name='var:name_api'])
    hadMember(var:Jupyter, var:libs)
    hadMember(var:libs, var:lib)

  endBundle
endDocument
```



<https://github.com/EnvriPlus-PROV/ProvTemplateCatalog>



Luc Moreau et al. A Templating System to Generate Provenance  
<https://eprints.soton.ac.uk/405025/1/provtemplate.pdf>

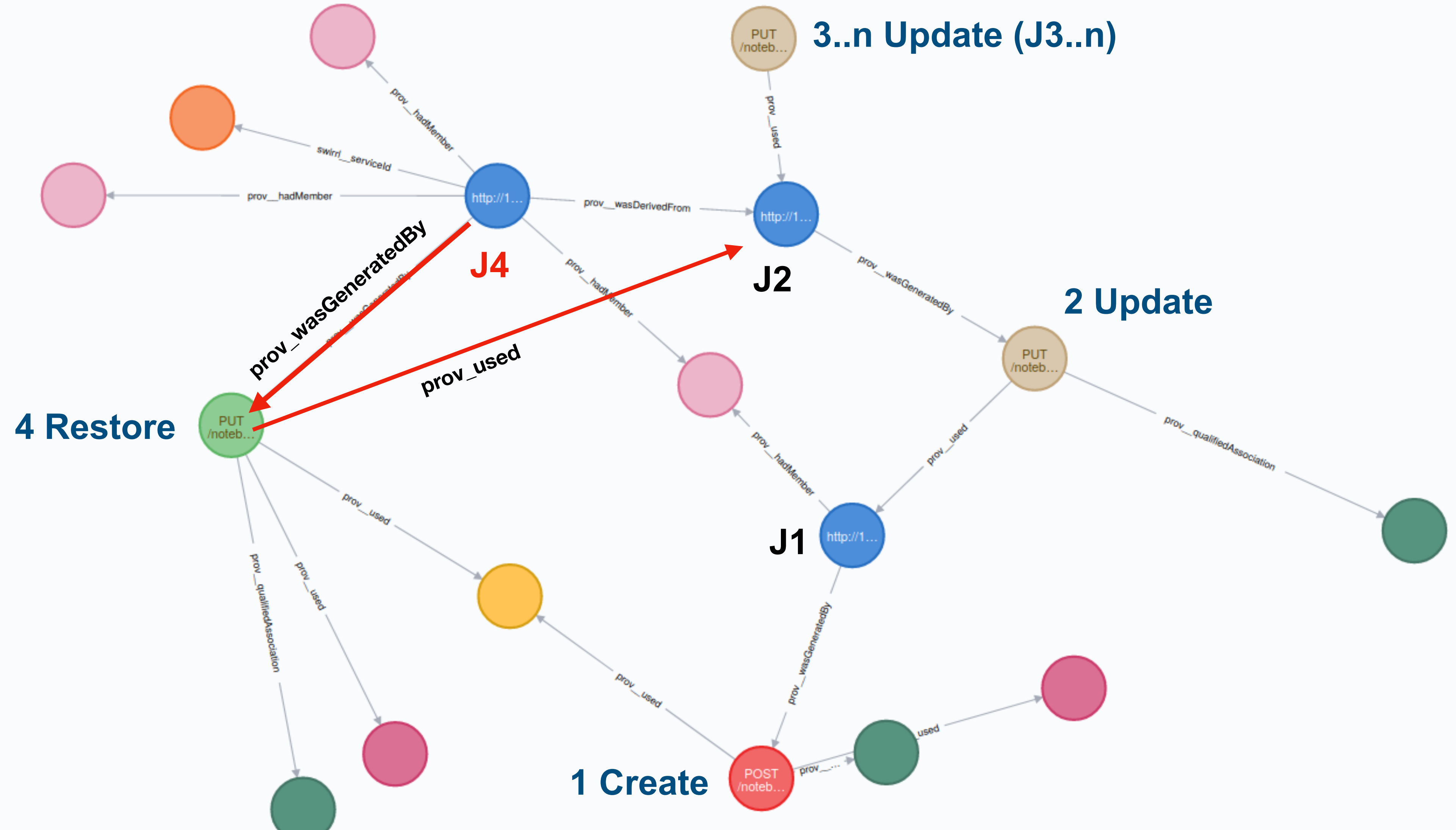
<https://openprovenance.org/store/documents/1968>

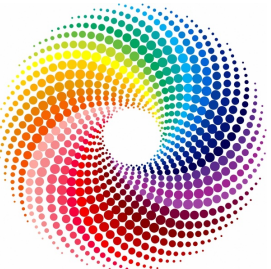


# Restoring Notebook Libraries (Prov Graph)



- \*(46)
- Resource(17)
- prov\_Activity(4)
- swirrl\_CreateNotebook(1)
- swirrl\_update(2)
- swirrl\_RestoreNotebook(1)
- prov\_Entity(9)
- swirrl\_SystemImage(1)
- prov\_Plan(2)
- \*(18)
- prov\_used(6)
- prov\_qualifiedAssociation(3)
- prov\_wasGeneratedBy(3)
- prov\_hadMember(4)
- prov\_wasDerivedFrom(1)
- swirrl\_serviceId(1)





## Provenance Storage, Acquisition and Query

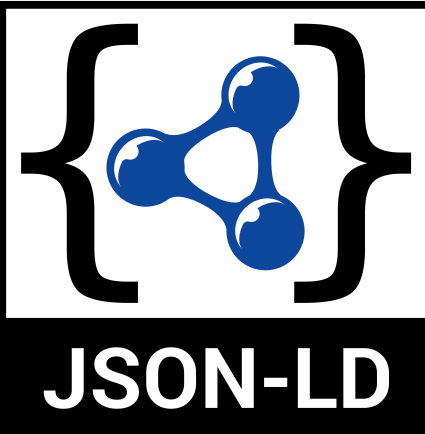
### provenance

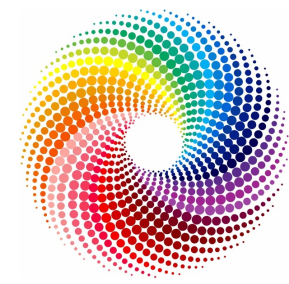


**POST** /**provenance** Saves provided provenance template expansion.

**GET** /**provenance/session/{sessionId}/activities** Returns the list of activities that are related to the session with their id's properties

**GET** /**provenance/activity/{activityId}** Returns the information about an activity based on the activityId





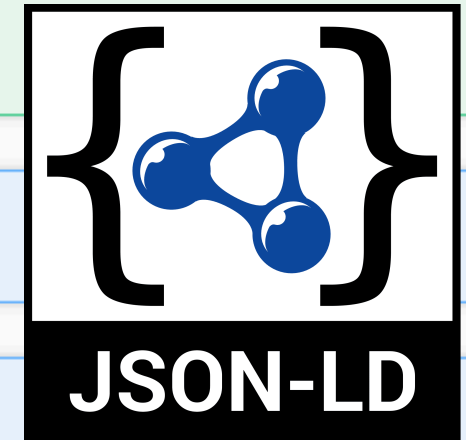
## Provenance Storage, Acquisition and Query

### provenance

**POST** `/provenance` Saves provided provenance template expansion.

**GET** `/provenance/session/{sessionId}/activities` Returns the list of activities that are related to the session with their id's properties

**GET** `/provenance/activity/{activityId}` Returns the information about an activity based on the activityId



↓

```
{ "@graph": [  
  { "swirrl:serviceId": "15c0e8a0-19f4-43ff-a619-b99845c01280",  
    "swirrl:sessionId": "a99f05e9-7570-46a3-b046-3c442ab1e23d",  
    "@type": [  
      "Resource",  
      "prov:Activity",  
      "swirrl:CreateNotebook"],  
    "prov:person": "Mock User",  
    "@id": "urn:uuid:375fe1aa-7957-4f19-81de-6549962fdbb4",  
    "prov:endedAtTime": "2020-12-01T14:32:05.000Z",  
    "prov:startedAtTime": "2020-12-01T14:31:21.337Z"},  
  { "swirrl:serviceId": "15c0e8a0-19f4-43ff-a619-b99845c01280",  
    "swirrl:sessionId": "a99f05e9-7570-46a3-b046-3c442ab1e23d",  
    "@type": [  
      "Resource",  
      "prov:Activity",  
      "swirrl:update"],  
    "prov:person": "Mock User",  
    "@id": "urn:uuid:daddea98-eb60-40d2-98e6-b2f026348da7",  
    "prov:endedAtTime": "2020-12-01T14:32:46.137Z",  
    "prov:startedAtTime": "2020-12-01T14:32:36.252Z"},  
], "@context": {...}}
```

# SWIRRL Jupyter Lab Extension



- Monitor Jobs

- Snapshot Controls

- Trace Activities and trigger rollback actions

File Edit View Run Kernel Tabs Settings Help

## SWIRRL

Notebook idle

Github  
aspinuso  
**SWITCH USER**  
Please review your access using [this link](#) to revoke your access tokens.

Snapshot  
Create a snapshot of your notebook and save it in your git repository.  
SU-Indicator **CREATE SNAPSHOT**  
Snapshot created: [repository url](#).

Activities  
Activity log of this notebook. Restore the notebook to a previous state.  
[LOAD ACTIVITIES](#)

Type	Created at	Action
Create	2020-11-18 16:59	<a href="#">RESTORE</a>
Update	2020-11-18 17:08	<a href="#">RESTORE</a>
Snapshot	2020-11-18 17:17	
Snapshot	2020-11-18 17:20	
Snapshot	2020-11-18 17:26	

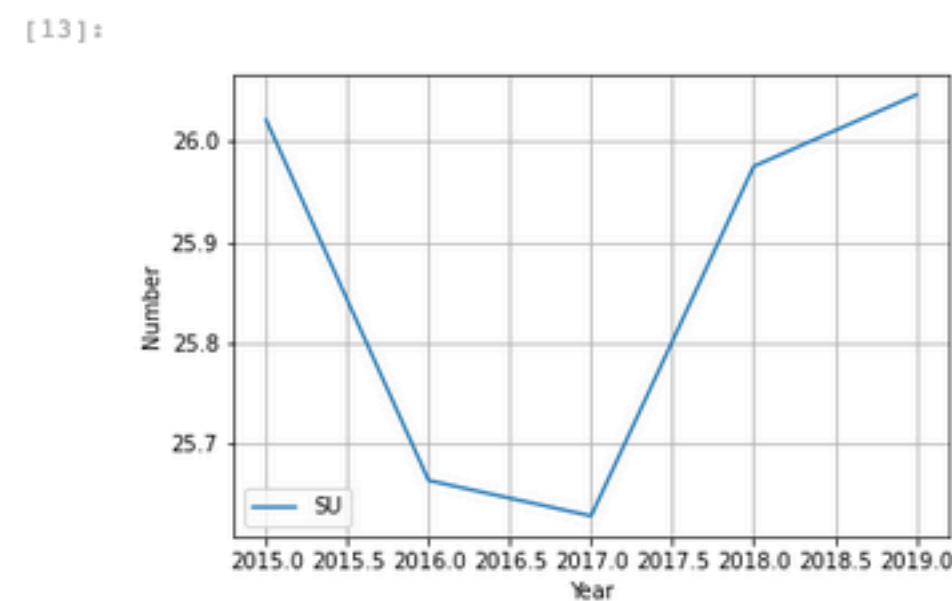
```
Test-ICCLIM-C4I.ipynb Terminal 1 Code
2020-11-18 17:33:03,933 *****
2020-11-18 17:33:06,908 Loading data: chunk 1/8 ...
2020-11-18 17:33:52,622 Loading data: chunk 2/8 ...
2020-11-18 17:34:42,806 Loading data: chunk 3/8 ...
2020-11-18 17:35:28,758 Loading data: chunk 4/8 ...
2020-11-18 17:35:39,935 Loading data: chunk 5/8 ...
2020-11-18 17:35:49,512 Loading data: chunk 6/8 ...
2020-11-18 17:36:18,272 Loading data: chunk 7/8 ...
2020-11-18 17:36:32,132 Loading data: chunk 8/8 ...
2020-11-18 17:36:41,195 *****
2020-11-18 17:36:41,196 *
2020-11-18 17:36:41,199 *          icclim                      V4.2.14      *
2020-11-18 17:36:41,201 *
2020-11-18 17:36:41,203 *
2020-11-18 17:36:41,203 *          Wed Nov 18 17:36:41 2020 GMT
2020-11-18 17:36:41,207 *
2020-11-18 17:36:41,207 *          END EXECUTION
2020-11-18 17:36:41,208 *
2020-11-18 17:36:41,209 *          CP SECS = 205.487
2020-11-18 17:36:41,210 *
2020-11-18 17:36:41,212 *****
```

Calculate spatial average

```
[12]: var = np.reshape(var, (var.shape[0], -1))
      result = np.mean(var, axis=1)
      print(result)
[26.02153  25.663391 25.628197 25.975288 26.046402]
```

Visualise the results

```
[13]: plt.figure()
      plt.plot(year_list, result, label='SU')
      plt.legend()
      plt.xlabel('Year')
      plt.ylabel('Number')
      plt.grid()
      name_fig = "su_icclim.png"
      plt.savefig("./"+name_fig)
      from IPython.display import Image
      Image(filename="./su_icclim.png")
```



# SWIRRL Jupyter Lab Extension



- Monitor Jobs

- Snapshot Controls

- Trace Activities and trigger rollback actions



# Other Resources and Demo



## **Binder Repository for Template Expansion Demo**

Used for KNMI internal Communication/Tutorial on integrating provenance in python scripts for climate studies processes

<https://gitlab.com/KNMI-OSS/swirrl/climate-scenarios-binder/-/blob/master/Template%20Expansion%20CS%20Single%20Calculation.ipynb>