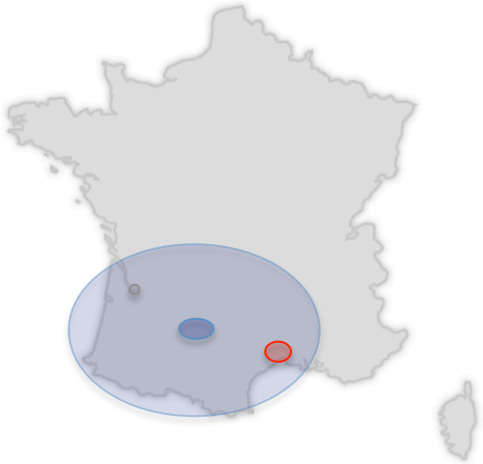
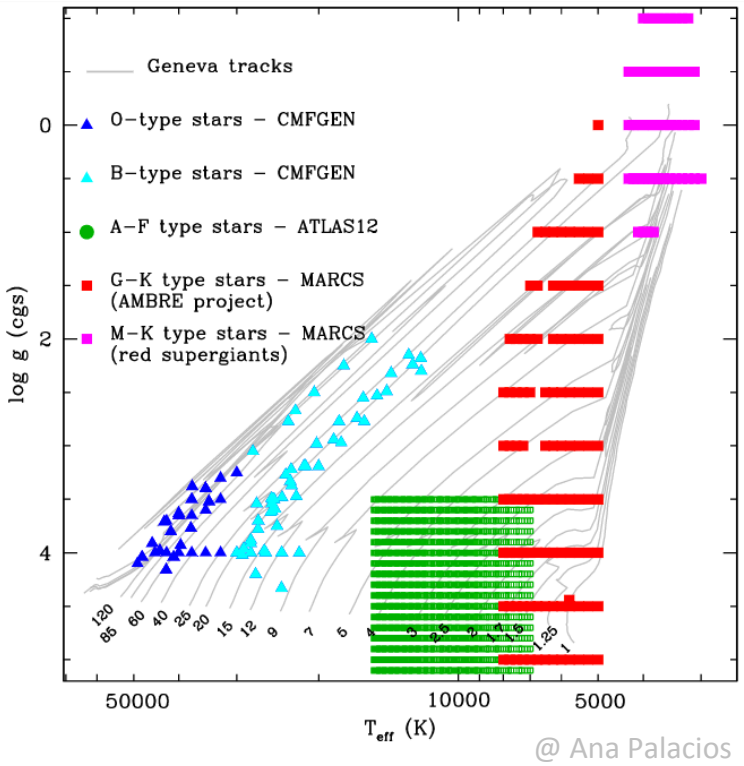
Pollux and Provenance

Scientists :
Ana Palacios, Agnès Lèbre
Software engineer :
Michèle Sanguillon



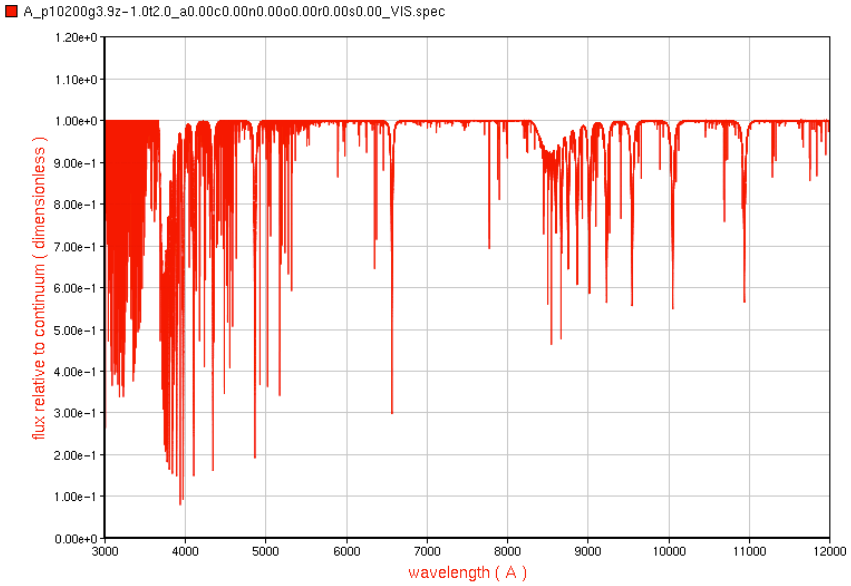
Pollux Database



@ Ana Palacios

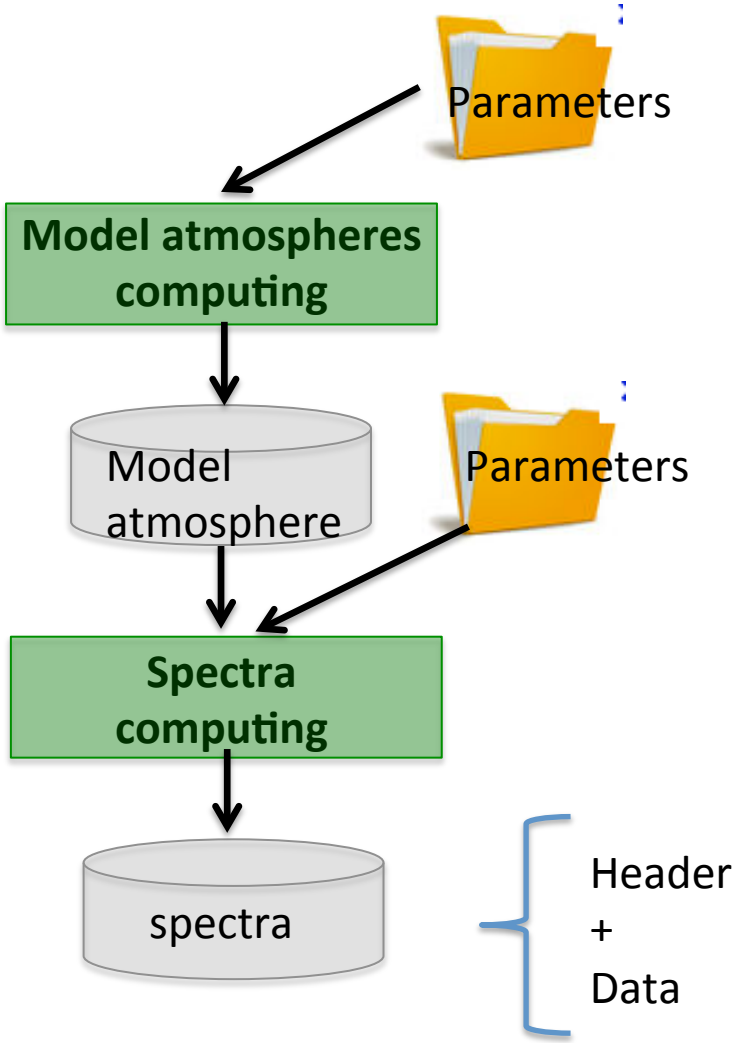
Database of very high resolution (R ≥ 150 000) synthetic spectra in the optical domain (3000 Å to 12000 Å).

Spectra exist for many spectral types (O to M and Wolf-Rayet stars.)



Data include the absolute flux and the flux normalized to the continuum.

Pollux data and workflow

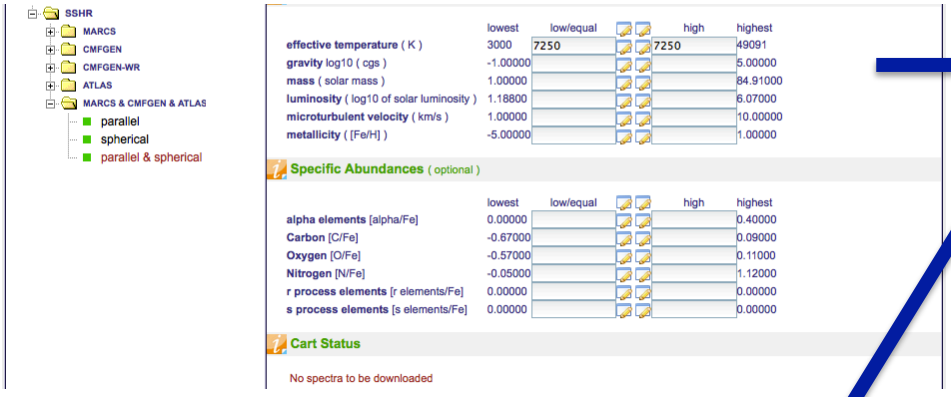


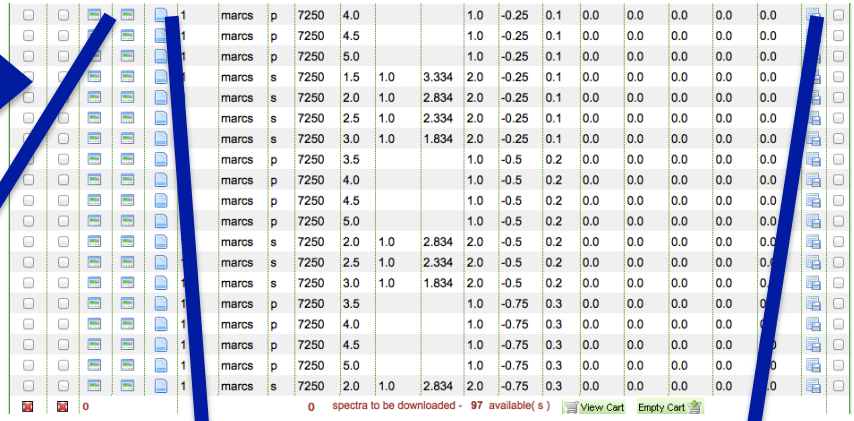
Only the spectra are available on the VO

Pollux accessibility

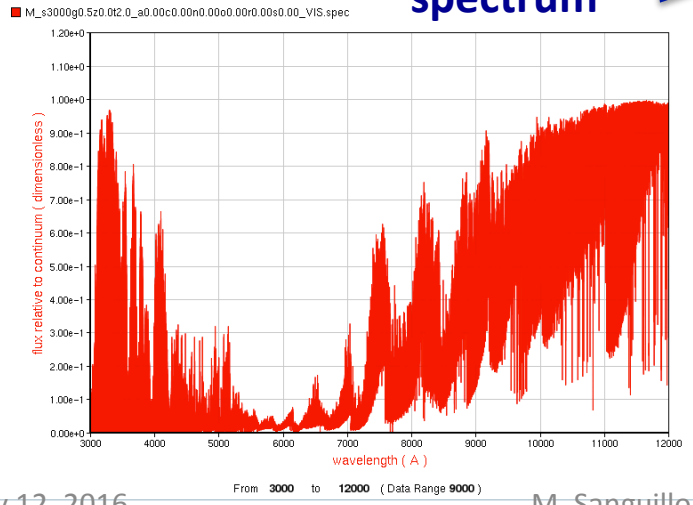
- via Web for Internet users:
  - Spectra in flat, fits, xml, votable formats (header + data)
- via SSA protocol for OV users:
  - Spectra in fits format (header + data)

Pollux accessibility via Web

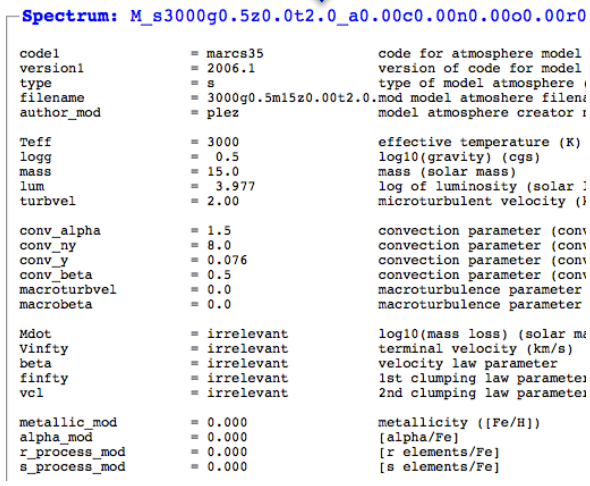




Display spectrum



Header

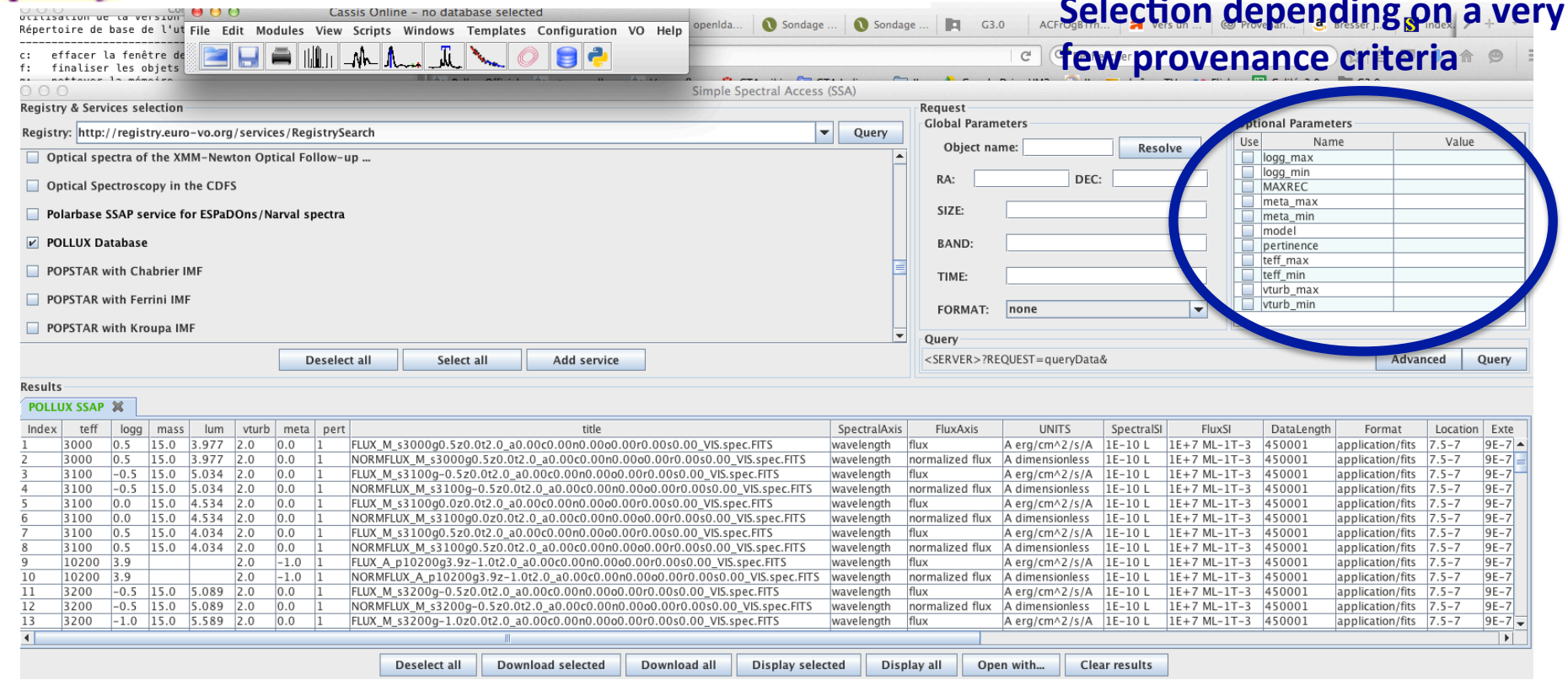


Download

Provenance information

Pollux accessibility via a VO tool

CASSIS
A free interactive spectrum analyser



Index	teff	logg	mass	lum	vturb	meta	pert	title	SpectralAxis	FluxAxis	UNITS	SpectralSI	FluxSI	DataLength	Format	Location	Exte
1	3000	0.5	15.0	3.977	2.0	0.0	1	FLUX_M_s3000g0.5z0.0t2.0_a0.00c0.00n0.00o0.00r0.00s0.00_VIS.spec.FITS	wavelength	flux	A erg/cm^2/s/A	1E-10 L	1E+7 ML-1T-3	450001	application/fits	7.5-7	9E-7
2	3000	0.5	15.0	3.977	2.0	0.0	1	NORMFLUX_M_s3000g0.5z0.0t2.0_a0.00c0.00n0.00o0.00r0.00s0.00_VIS.spec.FITS	wavelength	normalized flux	A dimensionless	1E-10 L	1E+7 ML-1T-3	450001	application/fits	7.5-7	9E-7
3	3100	-0.5	15.0	5.034	2.0	0.0	1	FLUX_M_s3100g-0.5z0.0t2.0_a0.00c0.00n0.00o0.00r0.00s0.00_VIS.spec.FITS	wavelength	flux	A erg/cm^2/s/A	1E-10 L	1E+7 ML-1T-3	450001	application/fits	7.5-7	9E-7
4	3100	-0.5	15.0	5.034	2.0	0.0	1	NORMFLUX_M_s3100g-0.5z0.0t2.0_a0.00c0.00n0.00o0.00r0.00s0.00_VIS.spec.FITS	wavelength	normalized flux	A dimensionless	1E-10 L	1E+7 ML-1T-3	450001	application/fits	7.5-7	9E-7
5	3100	0.0	15.0	4.534	2.0	0.0	1	FLUX_M_s3100g0.0z0.0t2.0_a0.00c0.00n0.00o0.00r0.00s0.00_VIS.spec.FITS	wavelength	flux	A erg/cm^2/s/A	1E-10 L	1E+7 ML-1T-3	450001	application/fits	7.5-7	9E-7
6	3100	0.0	15.0	4.534	2.0	0.0	1	NORMFLUX_M_s3100g0.0z0.0t2.0_a0.00c0.00n0.00o0.00r0.00s0.00_VIS.spec.FITS	wavelength	normalized flux	A dimensionless	1E-10 L	1E+7 ML-1T-3	450001	application/fits	7.5-7	9E-7
7	3100	0.5	15.0	4.034	2.0	0.0	1	FLUX_M_s3100g0.5z0.0t2.0_a0.00c0.00n0.00o0.00r0.00s0.00_VIS.spec.FITS	wavelength	flux	A erg/cm^2/s/A	1E-10 L	1E+7 ML-1T-3	450001	application/fits	7.5-7	9E-7
8	3100	0.5	15.0	4.034	2.0	0.0	1	NORMFLUX_M_s3100g0.5z0.0t2.0_a0.00c0.00n0.00o0.00r0.00s0.00_VIS.spec.FITS	wavelength	normalized flux	A dimensionless	1E-10 L	1E+7 ML-1T-3	450001	application/fits	7.5-7	9E-7
9	10200	3.9			2.0	-1.0	1	FLUX_A_p10200g3.9z-1.0t2.0_a0.00c0.00n0.00o0.00r0.00s0.00_VIS.spec.FITS	wavelength	flux	A erg/cm^2/s/A	1E-10 L	1E+7 ML-1T-3	450001	application/fits	7.5-7	9E-7
10	10200	3.9			2.0	-1.0	1	NORMFLUX_A_p10200g3.9z-1.0t2.0_a0.00c0.00n0.00o0.00r0.00s0.00_VIS.spec.FITS	wavelength	normalized flux	A dimensionless	1E-10 L	1E+7 ML-1T-3	450001	application/fits	7.5-7	9E-7
11	3200	-0.5	15.0	5.089	2.0	0.0	1	FLUX_M_s3200g-0.5z0.0t2.0_a0.00c0.00n0.00o0.00r0.00s0.00_VIS.spec.FITS	wavelength	flux	A erg/cm^2/s/A	1E-10 L	1E+7 ML-1T-3	450001	application/fits	7.5-7	9E-7
12	3200	-0.5	15.0	5.089	2.0	0.0	1	NORMFLUX_M_s3200g-0.5z0.0t2.0_a0.00c0.00n0.00o0.00r0.00s0.00_VIS.spec.FITS	wavelength	normalized flux	A dimensionless	1E-10 L	1E+7 ML-1T-3	450001	application/fits	7.5-7	9E-7
13	3200	-1.0	15.0	5.589	2.0	0.0	1	FLUX_M_s3200g-1.0z0.0t2.0_a0.00c0.00n0.00o0.00r0.00s0.00_VIS.spec.FITS	wavelength	flux	A erg/cm^2/s/A	1E-10 L	1E+7 ML-1T-3	450001	application/fits	7.5-7	9E-7

Selection depending on a very few provenance criteria

Currently no possibility to see the header or the provenance information because DATALINK is not yet implemented in spectra visualization tools .

Our motivations for Provenance

- **1) A Data Model to describe our data (spectra and provenance) => use of utypes for every piece of information:**

  - Spectrum DM 1.1 / Spectral DM 2.0:
    - adequate for spectra
    - but ours have 3 columns (wavelength, flux, normalized flux)
    - with a ObsConfig part
    - provenance information not included

  - SimDM:
    - adequate for simulations
    - complex
    - doesn't describe the spectrum

  - Future spectral DM(with 2 flux columns) + (ProvDM or SimDM):
    - could describe our spectra and the way they were produced

Our motivations for Provenance

2) A way to **display** all our provenance information in VO spectra tools, **easy readable** and **normalized**

Flat header (Web users)

```
header_name_SSHR = 'M_p7250g5.0z1.00t1.0_a0.00c0.00n0.00o0.00r0.00s0.00_VIS.spec.txt'
short_name_SSHR  = 'M_p7250g5.0z1.00t1.0_a0.00c0.00n0.00o0.00r0.00s0.00_VIS.spec'
Key_SSHR         = 'M_p7250g5.0z1.00t1.0_a0.00c0.00n0.00o0.00r0.00s0.00_conva1.5convny8.0convy0.076convb0.5mt(

code1           = 'marcs'                    / code for atmosphere model
version1        = '2008.5'                   / version of code for model atmosphere
type            = 'p'                        / type of model atmosphere (Spherical/Parallel)
filename        = 'p7250_g+5.0_m0.0_t01_st_z+1.00_a+0.00_c+0.00_n+0.00_o+0.00_r+0.00_s+0.00.mod'  / model atm(
author_mod      = 'Marcs-team'               / model atmosphere creator name

Teff            = '7250'                     / effective temperature (K)     - model atmosphere data
logg            = ' 5.0'                      / log10(gravity) (cgs)          - model atmosphere data
mass            = 'irrelevant'               / mass (solar mass)             - model atmosphere data
lum             = 'irrelevant'               / luminosity (solar luminosity)  - model atmosphere data
turbvel         = '1.00'                     / microturbulent velocity (km/s) - model atmosphere data

conv_alpha      = '1.5'                      / convection parameter (conva)   - model atmosphere data
conv_ny         = '8.0'                      / convection parameter (convny)  - model atmosphere data
conv_y          = '0.076'                    / convection parameter (convy)   - model atmosphere data
conv_beta       = '0.5'                      / convection parameter (convb)   - model atmosphere data
macroturbvel    = '0.0'                      / macroturbulence parameter (mt) - model atmosphere data
macrobeta       = '0.0'                      / macroturbulence parameter (mb) - model atmosphere data

Mdot            = 'irrelevant'               / log10(mass loss) (solar mass/year) - model atmosphere data
Vinfty          = 'irrelevant'               / terminal velocity (km/s)           - model atmosphere data
beta            = 'irrelevant'               / velocity law parameter             - model atmosphere data
finfty          = 'irrelevant'               / 1st clumping law parameter         - model atmosphere data
vcl             = 'irrelevant'               / 2nd clumping law parameter (km/s)  - model atmosphere data

metallic_mod    = '1.00'                    / metallicity ([Fe/H])
alpha_mod       = '0.000'                    / [alpha/Fe]
r_process_mod   = '0.000'                    / [r elements/Fe]
s_process_mod   = '0.000'                    / [s elements/Fe]
```

Easy to read and to understand

Keywords and format not normalized

No UCD, no utypes

Fits header (OV users)

```
<PARAM arraysize="60" datatype="char" name="short_name_SSHR" value="M_p7250g5.0z1.00t1.0_a0.00c0.00n0.00o
<PARAM arraysize="106" datatype="char" name="Key_SSHR" value="M_p7250g5.0z1.00t1.0_a0.00c0.00n0.00o0.00r0
<PARAM arraysize="5" datatype="char" name="code1" ucd="meta.id;meta.code" value="marcs">
<DESCRIPTION>code for atmosphere model</DESCRIPTION>
</PARAM>
<PARAM datatype="float" name="version1" ucd="meta.code;meta.version" value="2008.5">
<DESCRIPTION>version of code for model atmosphere</DESCRIPTION>
</PARAM>
<PARAM arraysize="1" datatype="char" name="type" ucd="meta.code.class" value="p">
<DESCRIPTION>type of model atmosphere (Spherical/Parallel)</DESCRIPTION>
</PARAM>
<PARAM arraysize="76" datatype="char" name="filename" ucd="meta.id;meta.file" value="p7250_g+5.0_m0.0_t01
<DESCRIPTION>model atmosphere filename</DESCRIPTION>
</PARAM>
<PARAM arraysize="10" datatype="char" name="author_mod" ucd="meta.bib.author" value="Marcs-team">
<DESCRIPTION>model atmosphere creator name</DESCRIPTION>
</PARAM>
<PARAM datatype="int" name="Teff" ucd="phys.temperature.effective" unit="K" value="7250">
<DESCRIPTION>effective temperature (K)     - model atmosphere data</DESCRIPTION>
</PARAM>
<PARAM datatype="float" name="logg" ucd="phys.gravity;arith.zp" unit="log(cm/s2)" value="5.0">
<DESCRIPTION>log10(gravity) (cgs)          - model atmosphere data</DESCRIPTION>
</PARAM>
<PARAM arraysize="10" datatype="char" name="mass" ucd="phys.mass" unit="M_sun" value="irrelevant">
<DESCRIPTION>mass (solar mass)             - model atmosphere data</DESCRIPTION>
</PARAM>
<PARAM arraysize="10" datatype="char" name="lum" ucd="phys.luminosity" unit="L_sun" value="irrelevant">
<DESCRIPTION>luminosity (solar luminosity)  - model atmosphere data</DESCRIPTION>
</PARAM>
```

Not very easy to read and to understand

VOTable (normalized)

UCD, no utypes

Our motivations for Provenance

3) Select spectra **on provenance criteria** as…

- **Use case 1:**

  Show me a list of synthetic spectra satisfying :
  - domain of wavelength = visible
  - domain of effective temperature = [4000, 5000]

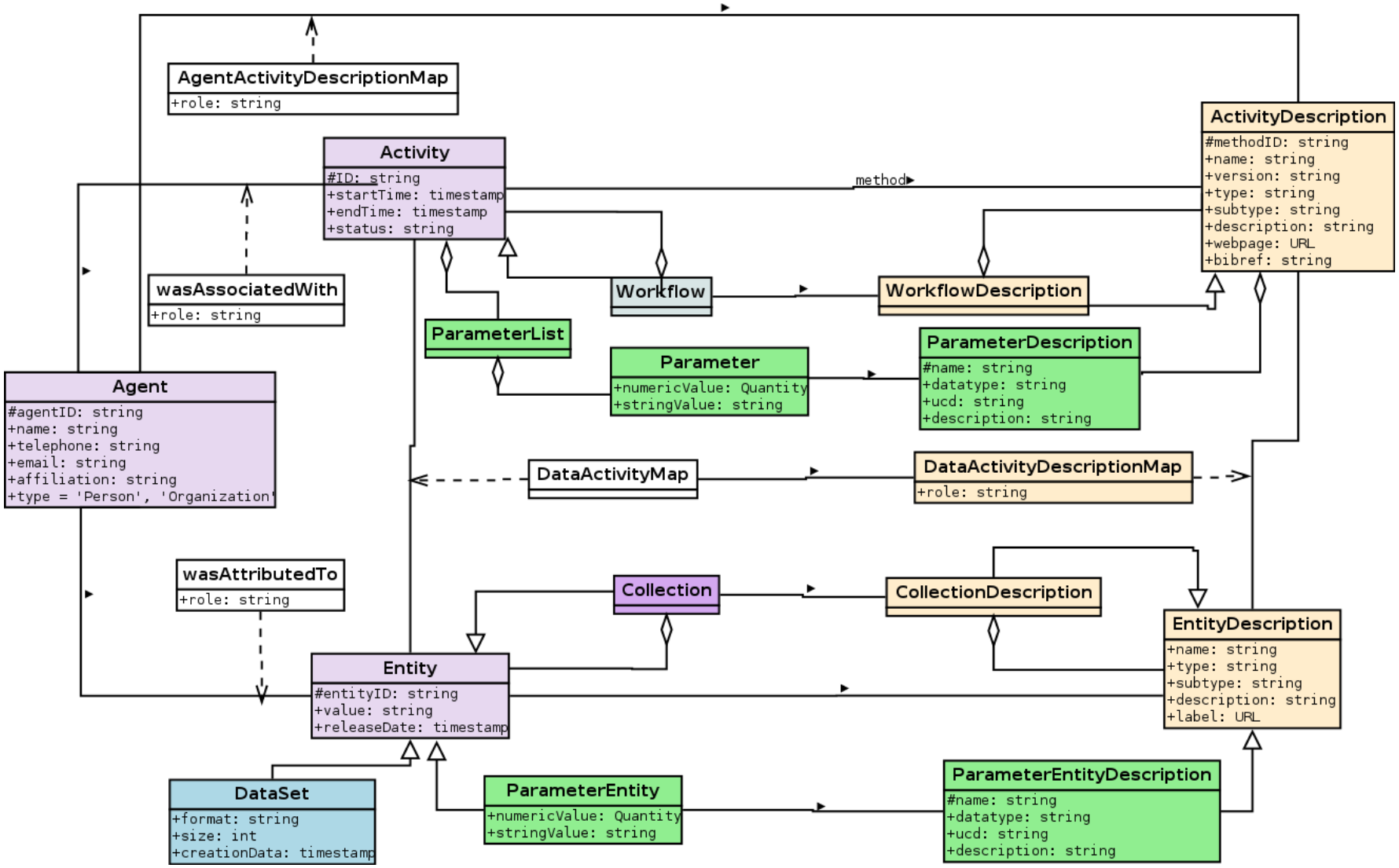
- **Use case 2:**

  Show me a list of synthetic spectra satisfying :
  - code for model atmosphere = MARCS
  - type of model atmosphere = spherical

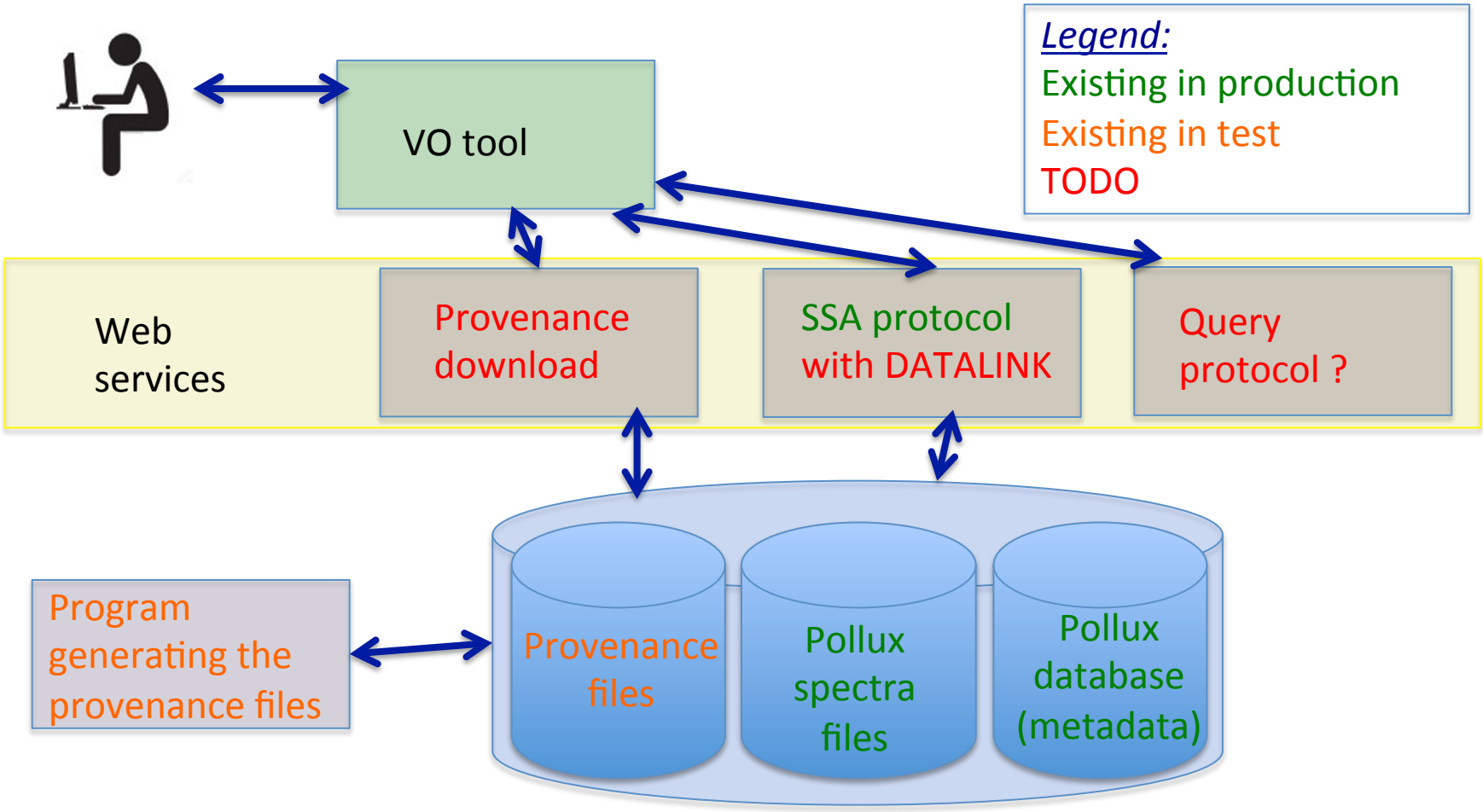
- **Use case 3:**

  Show me a list of synthetic spectra satisfying :
  - code for spectral synthesis = turbospectrum
  - version of this code = 2008.1

Data Model used



A beginning of implementation



Creation of provenance files

- Code : use of the prov 1.4.0 python package (developed at the University of Southampton)
  - A library for W3C Provenance Data Model
  - Can generate different formats:
    - Serialized formats: PROVN, JSON, XML, VOTABLE
    - Graphic formats: PNG, SVG, PDF
  - Uses namespaces:
    - prov: http://www.w3.org/ns/prov#
    - voprov (TBD): http://www.ivoa.net/documents/dm/provdm/voprov/
    - polluxData: http://dev-pollux/datalink/provenance?Id=

- Validation: use of the **Southampton Provenance Suite** https://provenance.ecs.soton.ac.uk/

Creation of provenance files

- **Serialized formats:**
  - JSON: 3906 lines
  - PROVN : 613 lines
  - XML: 3800 lines
  - VOTable : not yet implemented

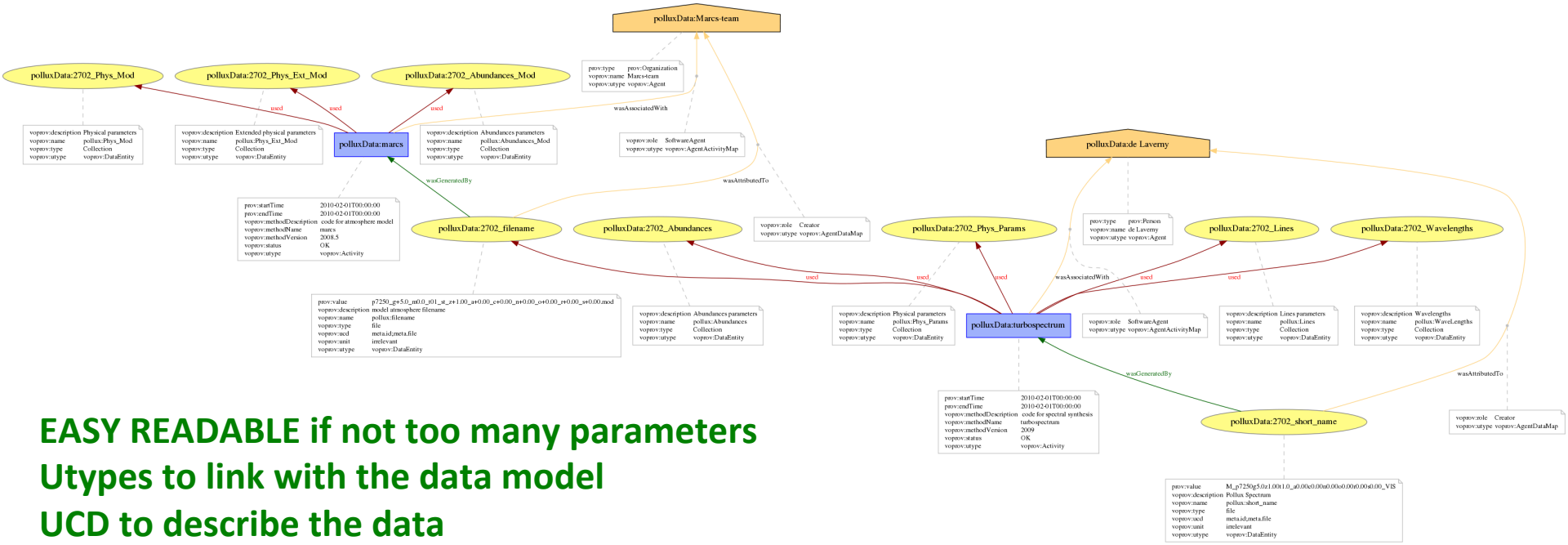
```
"entity": {
"pollux:14800g4.1z0.0a0.0C0.0.mod_2012_Teff": {
    "voprov:ucd": "phys.temperature.effective",
    "voprov:type": "int",
    "voprov:description": "effective temperature (K) ",
    "voprov:unit": "K",
    "prov:value": "14800"
  },
```

```
<prov:entity prov:id="polluxData:2702_Teff">
  <prov:value xsi:type="xsd:string">3000</prov:value>
  <voprov:description>effective temperature (K)</voprov:description>
  <voprov:name>pollux:Teff</voprov:name>
  <voprov:type>int</voprov:type>
  <voprov:ucd>phys.temperature.effective</voprov:ucd>
  <voprov:unit>K</voprov:unit>
  <voprov:utype>voprov:DataEntity</voprov:utype>
 </prov:entity>
```

```
entity(pollux:14800g4.1z0.0a0.0C0.0.mod_2012_Teff, [voprov:description="effective temperature (K)    - model atmosphere data", voprov:type="int", prov:value="14800", voprov:unit="K", voprov:ucd="phys.temperature.effective"])
```

Creation of provenance files

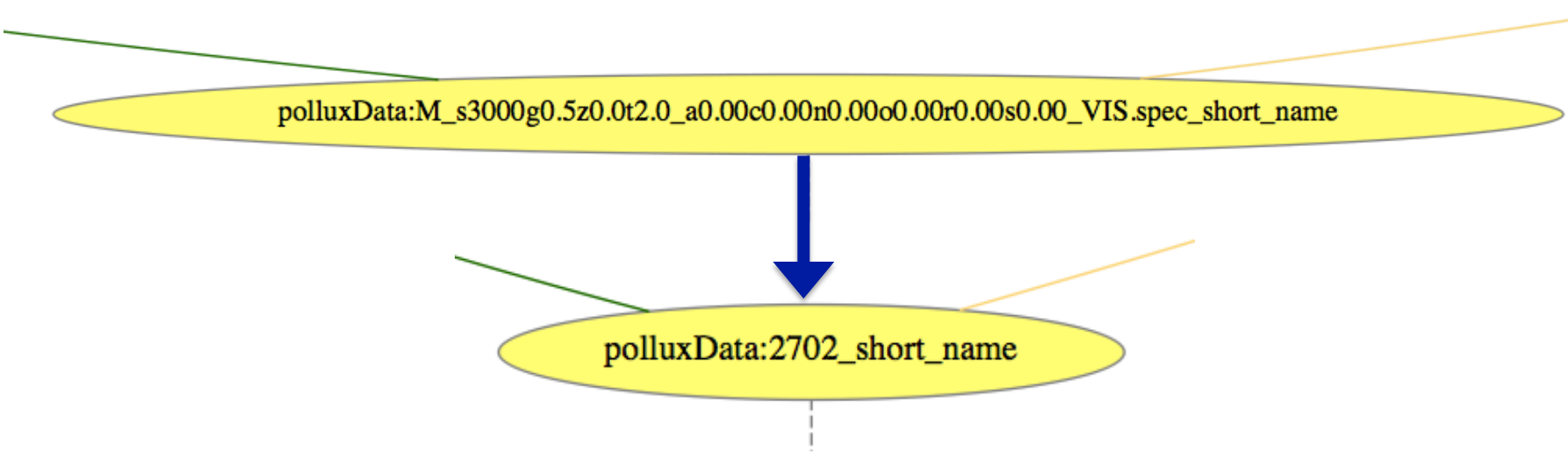
- **Graphic formats:**
  - SVG: you can click on each declaration
  - PNG : only a picture
  - PDF : currently a picture



EASY READABLE if not too many parameters
Utypes to link with the data model
UCD to describe the data

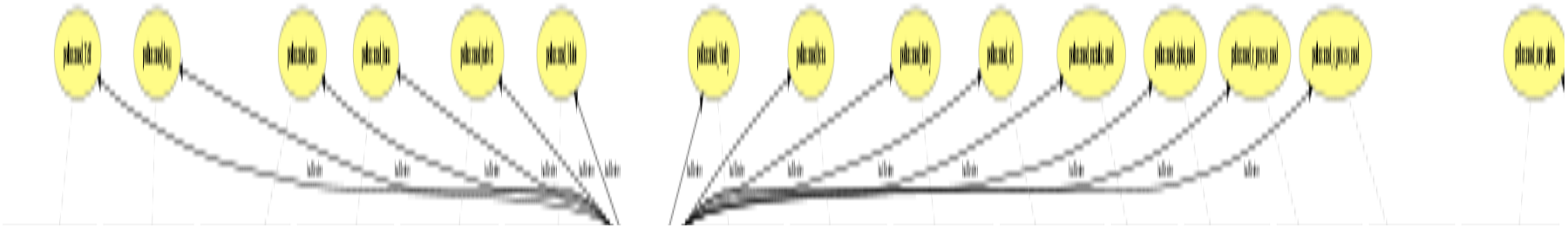
Creation of provenance files
Problems encountered

- Name of instances too long (Pollux key too long)
  => graphics not easy to read
  Solution: add a column in the database to put a number

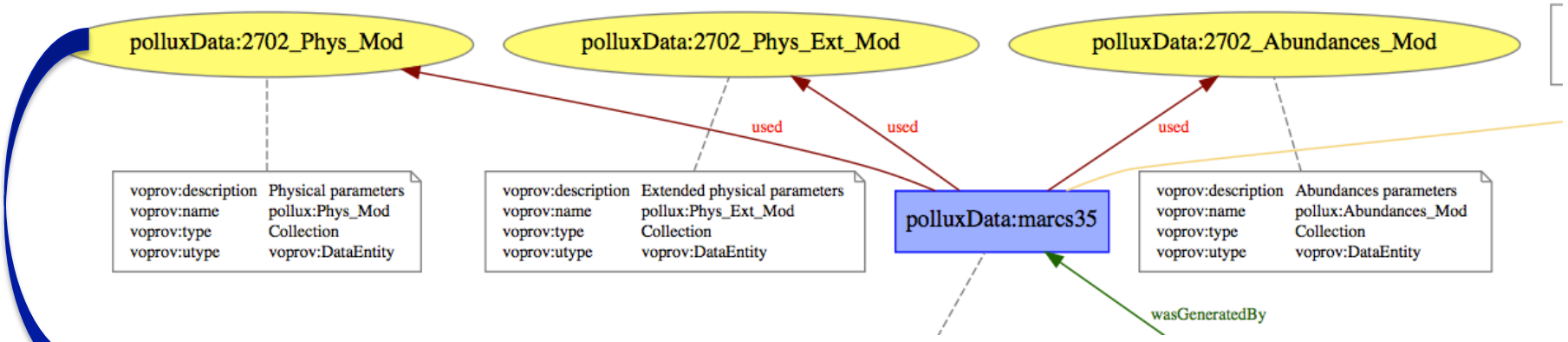


Creation of provenance files
Problems encountered

- **Graphic too big so unreadable if too many input parameters**



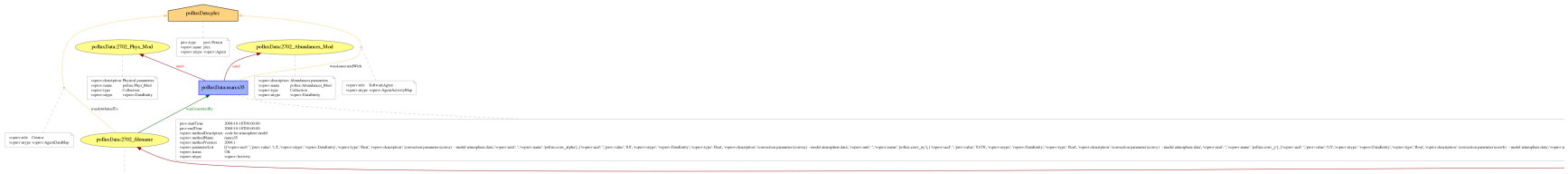
Solution: input parameters grouped



http://dev-pollux/datalink/provenance?Id=2702_Phys_Mod

Creation of provenance files
Problems encountered

- **Activity definition line too long in the Parameters directly attached to the activity**



Solution: we have to modify the code to display the parameters of the activity in an other way (several lines, not the whole information)

Conclusions

- It is just a beginning of implementation
- Interesting stuffs:
  - Existing tools:
  - Prov python package + Southampton suite for validation
  - SVG format with the use of namespaces
- A lot of things to do:
  - Explore SimDM
  - Implement DataLink descriptors and link resources
  - Serialize the provenance in a VOTable
  - Think about how to query