Provenance-enabled Automatic Data Publication

James Frew, Greg Janée, and Peter Slaughter
Earth Research Institute, University of California, Santa Barbara

From The Times

October 13, 2009

Unrecognised Leonardo da Vinci portrait revealed by his fingerprint

Provenance: motivation

- At the toolbar (menu, whatever) associated with a document there is a button marked "Oh, yeah?". You press it when you lose that feeling of trust. It says to the Web, "so how do I know I can trust this information?"

...

The result of pressing on the "Oh, yeah?" button is either a list of assumptions on which the trust is based, or of course an error message indicating either that a signature has failed, or that the system couldn't find a path of trust from you to the page.

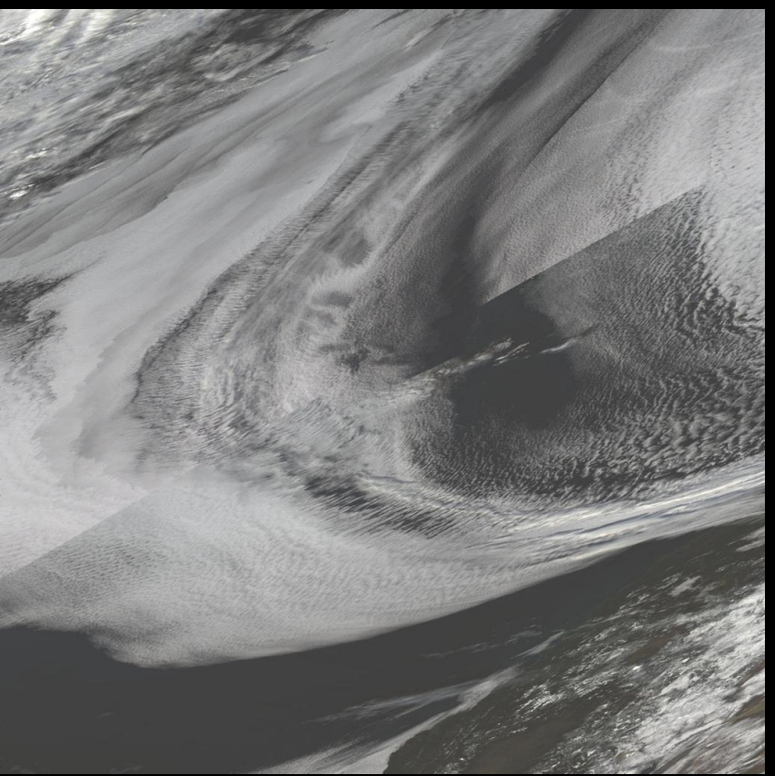
- —Tim Berners-Lee (1995)

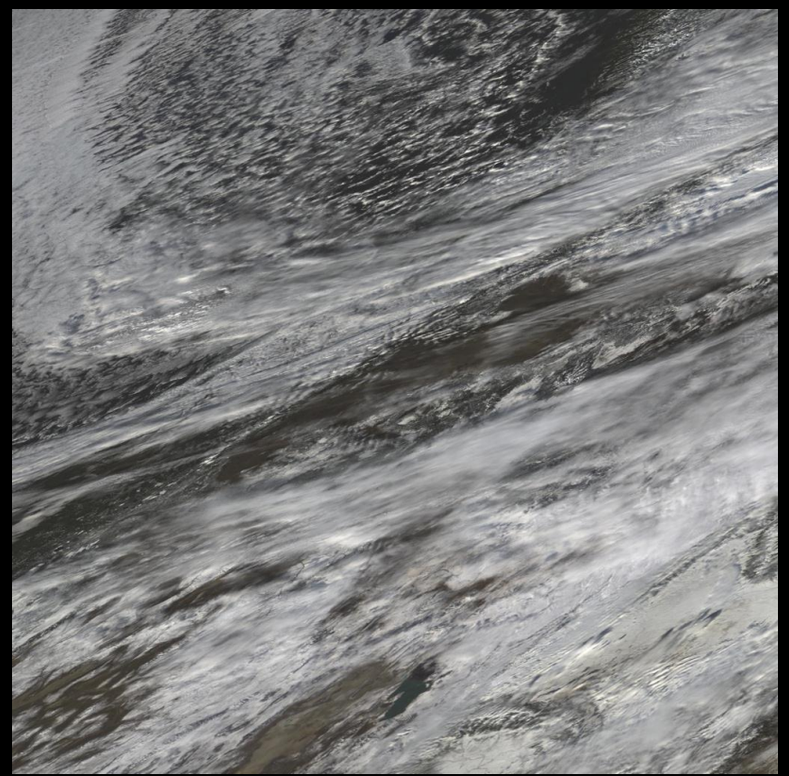
Provenance: “working’ definition

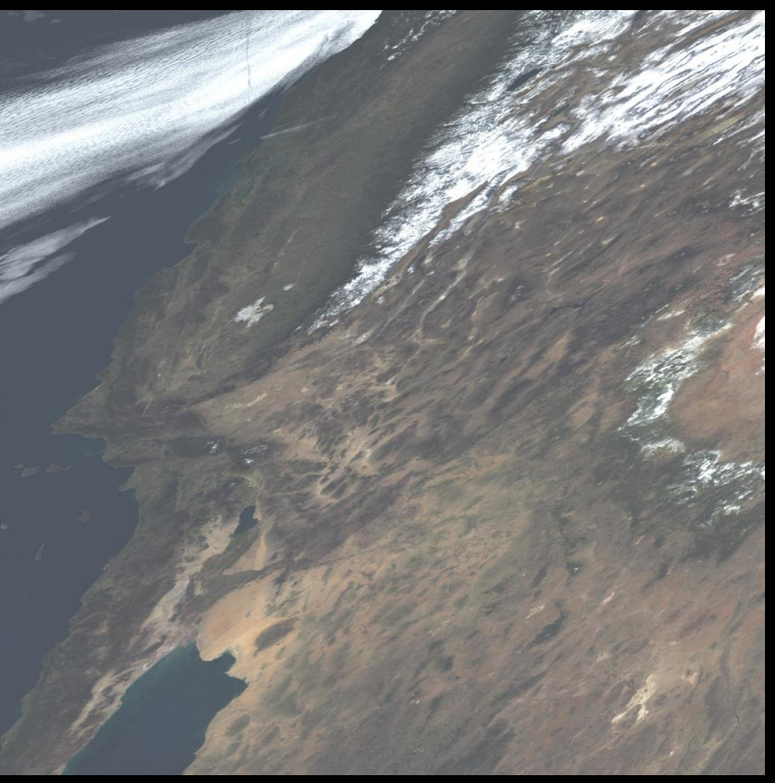
- **Provenance of a resource is a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource.** Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility. Provenance assertions are a form of contextual metadata and can themselves become important records with their own provenance.

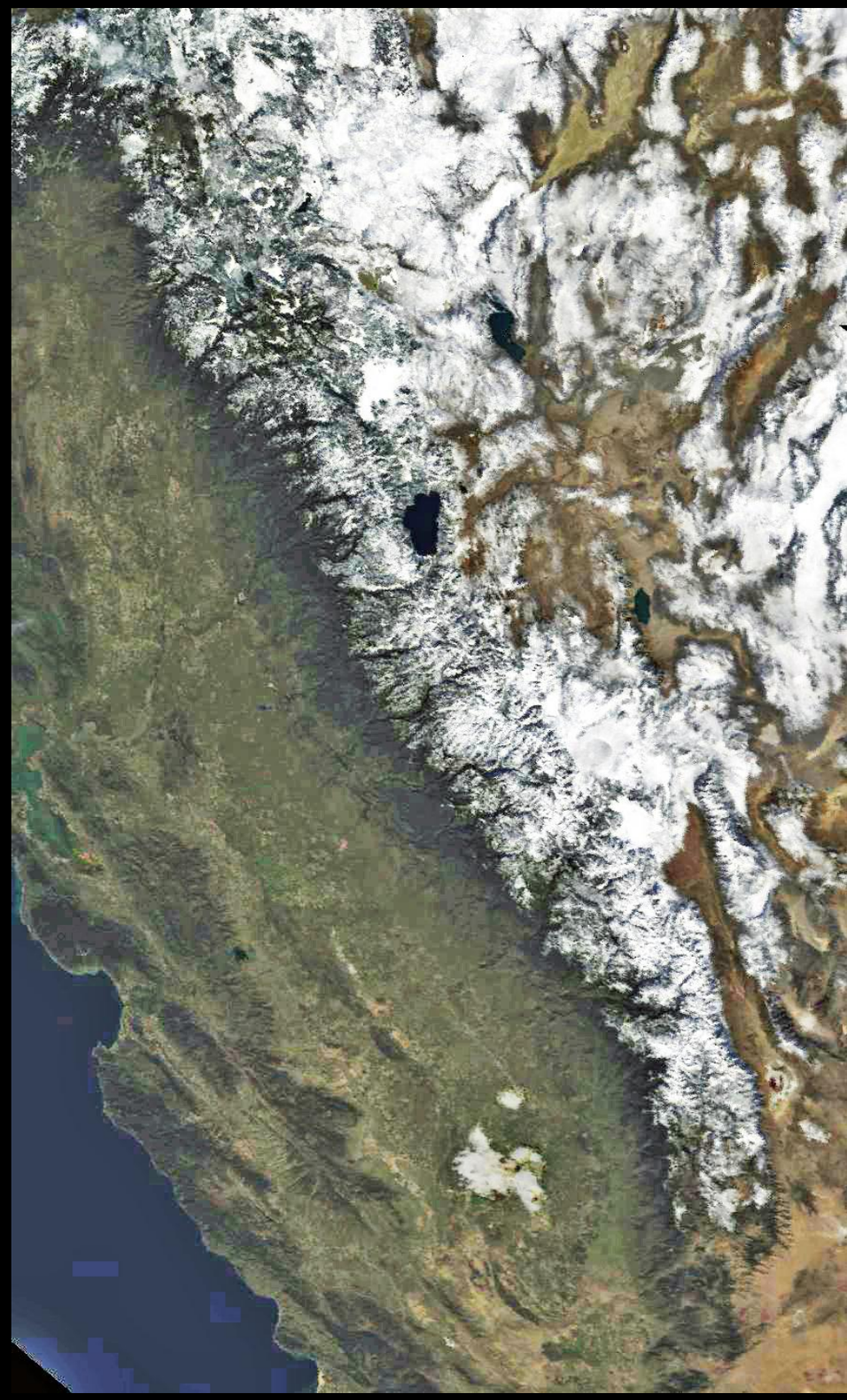
—W3C Provenance Incubator Group (2010)

from 50 Kft
to
0.5 ft









mosaic.sh:

```
  mosaicFn="MOD09GA.A2008019.sn.005.hdf"  mrtmosaic -i
tile.lis -o $mosaicFn  resample -p MRT.prm -g MRT.log
```

tile.lis:

```
  MOD09GA.A2008019.h08v04.005.2008022125449.hdf
MOD09GA.A2008019.h08v05.005.2008022134646.hdf
MOD09GA.A2008019.h09v04.005.2008022151755.hdf
```

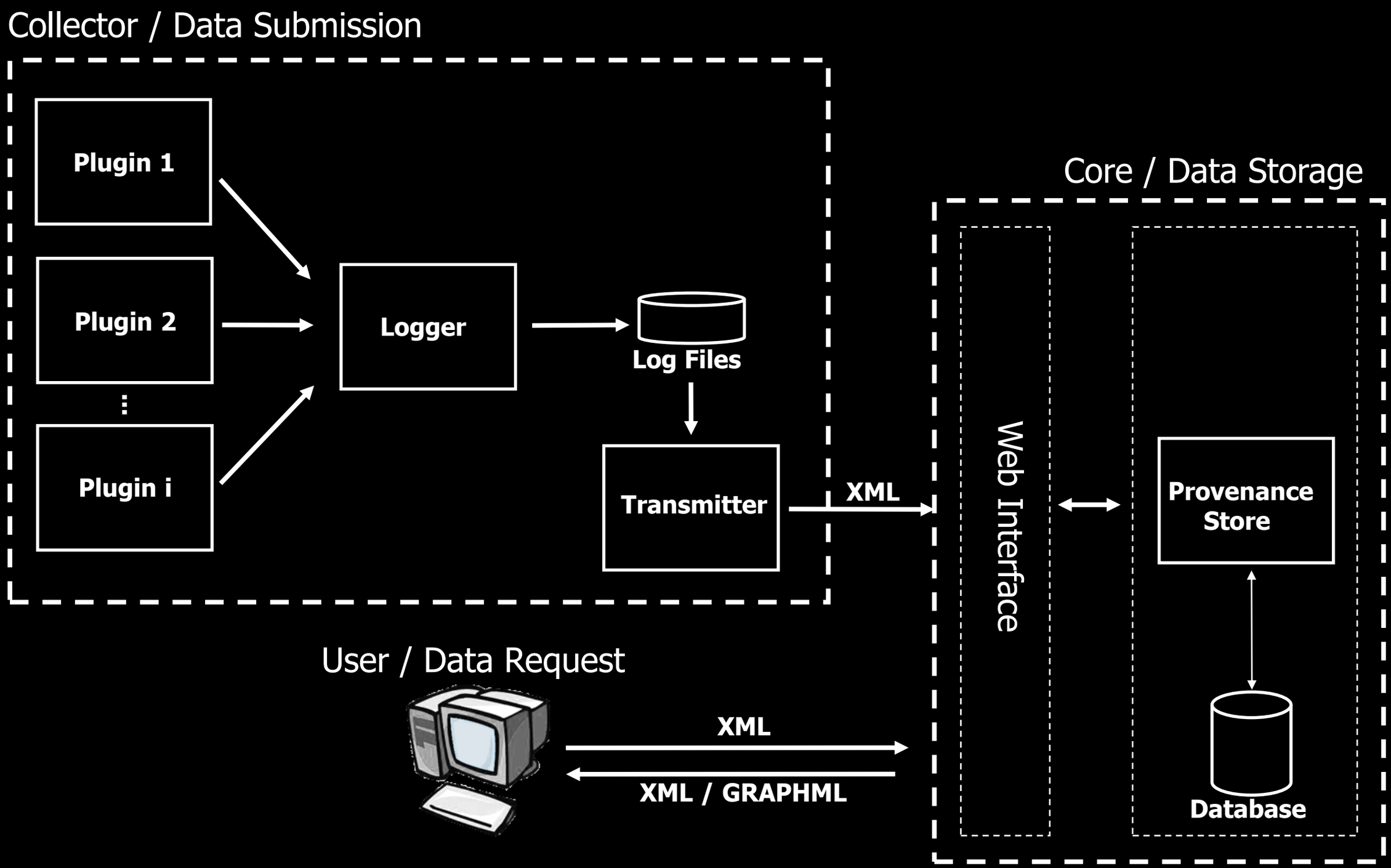
MRT.prm:

```
  INPUT_FILENAME=./MOD09GA.A2008019.sn.005.hdf
SPATIAL_SUBSET_TYPE=INPUT_LAT_LONG
SPATIAL_SUBSET_UL_CORNER=(41.5000 -122.4000)
SPATIAL_SUBSET_LR_CORNER=(35.0000 -117.6000)
OUTPUT_FILENAME=MOD09GA.A2008019.sn_cal-aea.005.Refl.hdf
RESAMPLING_TYPE=NN  OUTPUT_PROJECTION_TYPE=AEA  DATUM=WGS84
OUTPUT_PROJECTION_PARAMETERS=(0.0 0.0 34.00 40.50 -120.00 \
    0.00 0.00 -4000000.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00)
OUTPUT_PIXEL_SIZE=500  SPECTRAL_SUBSET=(0 0 0 0 0 0 0 0 0 0 1
1 1 1 1 1 1 0 0 0)
```

provenance in ES3

- input file(s) → process → output file(s)
- collected automatically by tracing
  - process creation
  - program execution
  - filesystem I/O

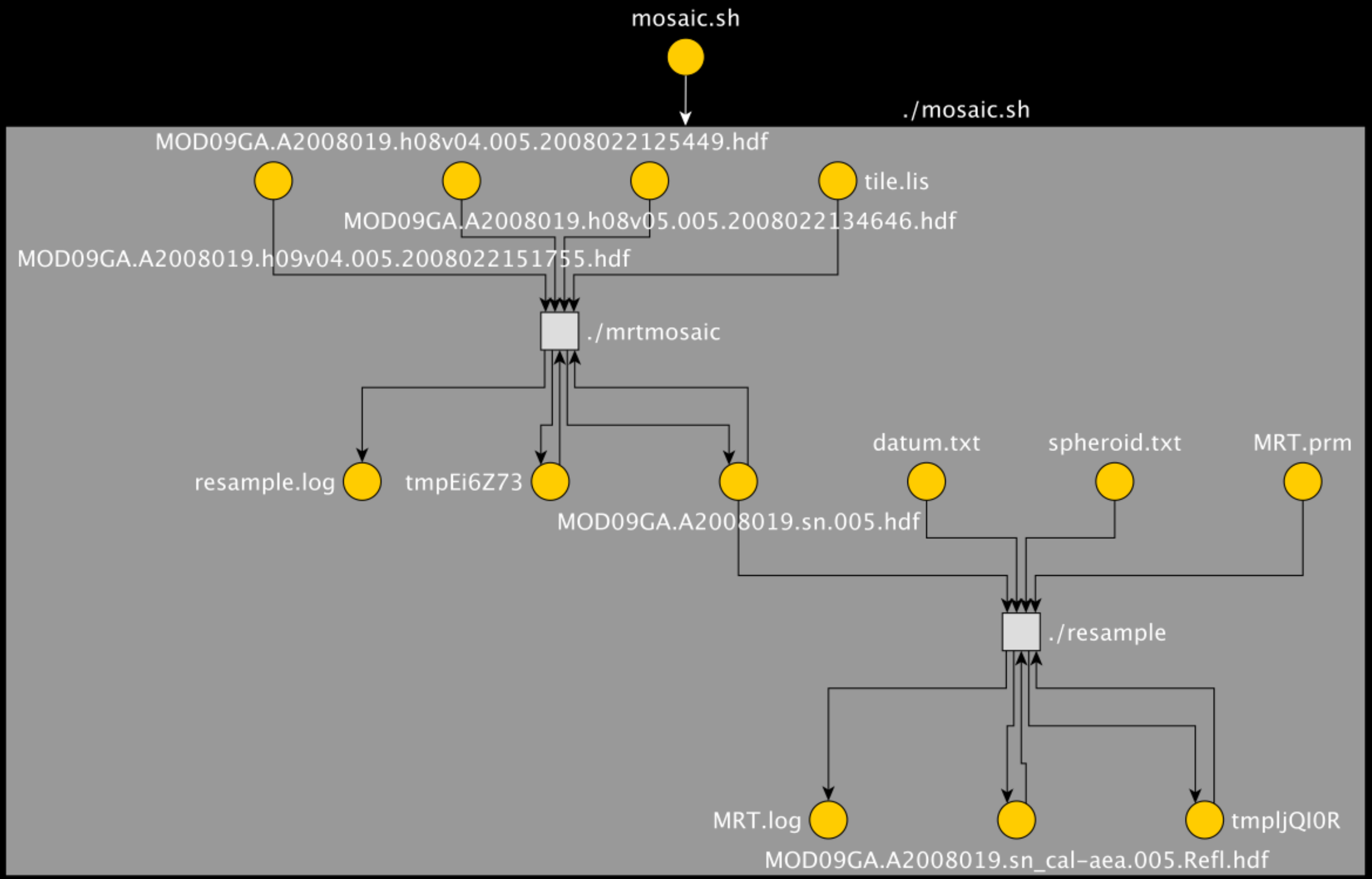
ES3 architecture

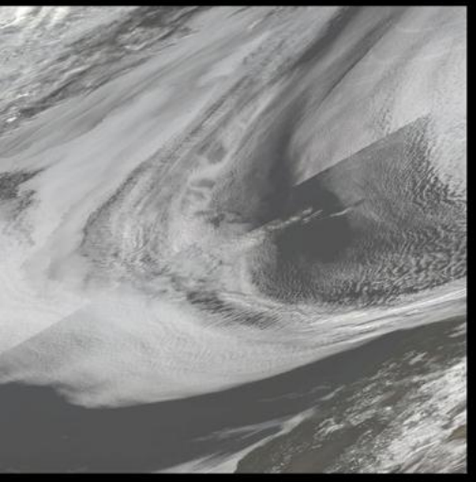


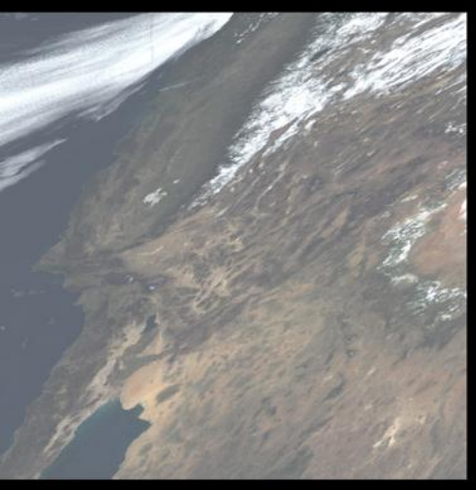
```
4810  1213121515.708913 execve("./mosaic.sh", ["mosaic.sh"], [...]) = 0...4810
1213121515.712317 open("/lib/libc.so.6", O_RDONLY) = 3...4810  1213121515.717415
open("./mosaic.sh", O_RDONLY|O_LARGEFILE) = 34810  1213121520.732852 clone(...) = 48304830
1213121520.735487 execve("./mrtmosaic", \
         ["mrtmosaic", "-i", "tile.lis", "-o", "MOD09GA.A2008019.sn.005.hdf"], [...]) = 04830
1213121520.768912 open("tmpEi6Z73", O_WRONLY|O_CREAT|O_TRUNC, 0666) = 34830  1213121520.769899
open("tile.lis", O_RDONLY) = 34830  1213121521.159965
open("MOD09GA.A2008019.h08v04.005.2008022125449.hdf", O_RDONLY) = 44830  1213121521.290125
open("MOD09GA.A2008019.h08v05.005.2008022134646.hdf", O_RDONLY) = 44830  1213121521.715161
open("MOD09GA.A2008019.h09v04.005.2008022151755.hdf", O_RDONLY) = 44830  1213121689.009340
open("tmpEi6Z73", O_WRONLY|O_CREAT|O_APPEND, 0666) = 44830  1213121522.161875
open("MOD09GA.A2008019.sn.005.hdf", O_RDWR|O_CREAT|O_TRUNC, 0666) = 44830  1213121689.010752
open("resample.log", O_WRONLY|O_CREAT|O_APPEND, 0666) = 44830  1213121689.071644 exit_group(0)
= ?4810  1213121689.299345 clone(...) = 49044904  1213121689.301804 execve("./resample", \
         ["resample", "-p", "MRT.prm", "-g", "MRT.log"], [...]) = 04904  1213121689.654760
open("tmpljQI0R", O_WRONLY|O_CREAT|O_APPEND, 0666) = 34904  1213121689.657942 open("MRT.prm",
O_RDONLY) = 34904  1213121689.864752 open("./MOD09GA.A2008019.sn.005.hdf", O_RDONLY) = 34904
1213121690.623884 open("MOD09GA.A2008019.sn_cal-aea.005.Refl.hdf", \
         O_RDWR|O_CREAT|O_TRUNC, 0666) = 44904  1213121714.410092 open("MRT.log",
O_WRONLY|O_CREAT|O_APPEND, 0666) = 34904  1213121714.457637 open("MOD09GA.A2008019.sn_cal-
aea.005.Refl.hdf", O_RDONLY) = 34904  1213121714.458947 open("MOD09GA.A2008019.sn_cal-
aea.005.Refl.hdf", O_RDWR) = 34904  1213121714.463607 open("./MOD09GA.A2008019.sn.005.hdf",
O_RDONLY) = 44810  1213121714.615284 exit_group(0)   = ?
```

```
<init time="20080610T181514Z" stime="20080610T181155.707233Z"
      pstime="20080610T181155.707233Z" pid="4783" ppid="4783" language="bash"
      user="peter" hostname="localhost.localdomain"></init><exec
time="20080610T181515Z" routine="./mosaic.sh" pid="4810">  <arguments>  </arguments>
<io>    <pipe read="true" id="std-in"/>    <pipe write="true" id="std-out"/>
<pipe write="true" id="std-err"/>    <file read="true">/etc/ld.so.cache</file>
<file read="true">/lib/libtermcap.so.2</file>    <file
read="true">/lib/libdl.so.2</file>    <file read="true">/lib/libc.so.6</file>
<file read="true" write="true">/dev/tty</file>    <file
read="true">/usr/lib/locale/locale-archive</file>    <file
read="true">/proc/meminfo</file>    <file read="true">/usr/lib/gconv/gconv-
modules.cache</file>    <file
read="true">/home/peter/Test/ES3/RegressionTests/MODSCAG/mosaic.sh</file>
</io></exec>
```

```
<ES3Request type="storeTransformation">  <transformation>
<timestamp type="execution">20080610T181515Z</timestamp>
<provenance>      <link>        <type>1/0</type>
<fromUuid>
          7af82a69-fa7a-4aec-abdf-eb009f5e2cab
        </fromUuid>      </link>    </provenance>
<collection>/default</collection>    <workflowUuid>
      b2189b33-349c-434d-bf73-3f8817dccbd5
    </workflowUuid>    <containsWorkflowUuid>
      2c4310db-4949-4fab-a82e-1282432257c3
    </containsWorkflowUuid>    <uuid>197dc9ee-3dbf-447b-
871a-e11a0288a7ba</uuid>    <name>./mosaic.sh</name>
</transformation></ES3Request>
```


mosaic.sh
./mosaic.sh
MOD09GA.A2008019.h08v04.005.2008022125449.hdf
tile.lis
MOD09GA.A2008019.h08v05.005.2008022134646.hdf
MOD09GA.A2008019.h09v04.005.2008022151755.hdf
./mrtmosaic
resample.log
tmpEi6Z73
datum.txt
spheroid.txt
MRT.prm
MOD09GA.A2008019.sn.005.hdf
./resample
MRT.log
tmpljQI0R
MOD09GA.A2008019.sn_cal-aea.005.Refl.hdf





mosaic.sh:

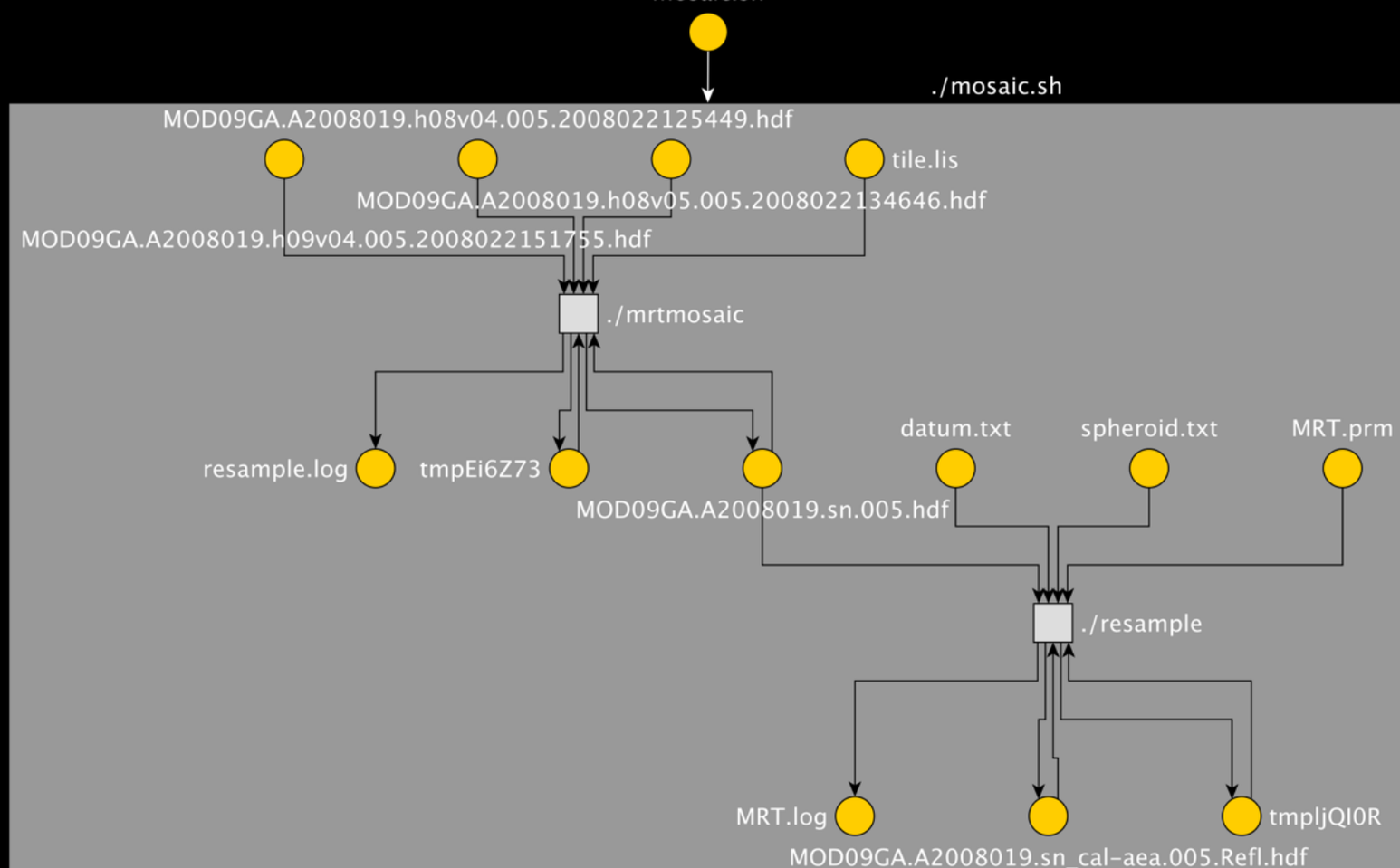
```
mosaicFn="MOD09GA.A2008019.sn.005.hdf"
mrtmosaic -i tile.lis -o $mosaicFn
resample -p MRT.prm -g MRT.log
```

tile.lis:

```
MOD09GA.A2008019.h08v04.005.2008022125449.hdf
MOD09GA.A2008019.h08v05.005.2008022134646.hdf
MOD09GA.A2008019.h09v04.005.2008022151755.hdf
```

MRT.prm:

```
INPUT_FILENAME=.
SPATIAL_SUBSET_T
SPATIAL_SUBSET_U
SPATIAL_SUBSET_L
OUTPUT_FILENAME=
RESAMPLING_TYPE=
OUTPUT_PROJECTIO
DATUM=WGS84
OUTPUT_PROJECTIO
  0.00 0.00 -400
OUTPUT_PIXEL_SIZ
SPECTRAL_SUBSET=
```



data publication

- evaluate object's antecedents against publication assertions
- if antecedents justify publication, then object is publishable

“publish” tool

- retrieve object’s provenance
- traverse depth-first
- foreach antecedent
  - automatically endorse if assertion valid
  - else manually endorse
- save endorsements in provenance graph

automatic endorsement

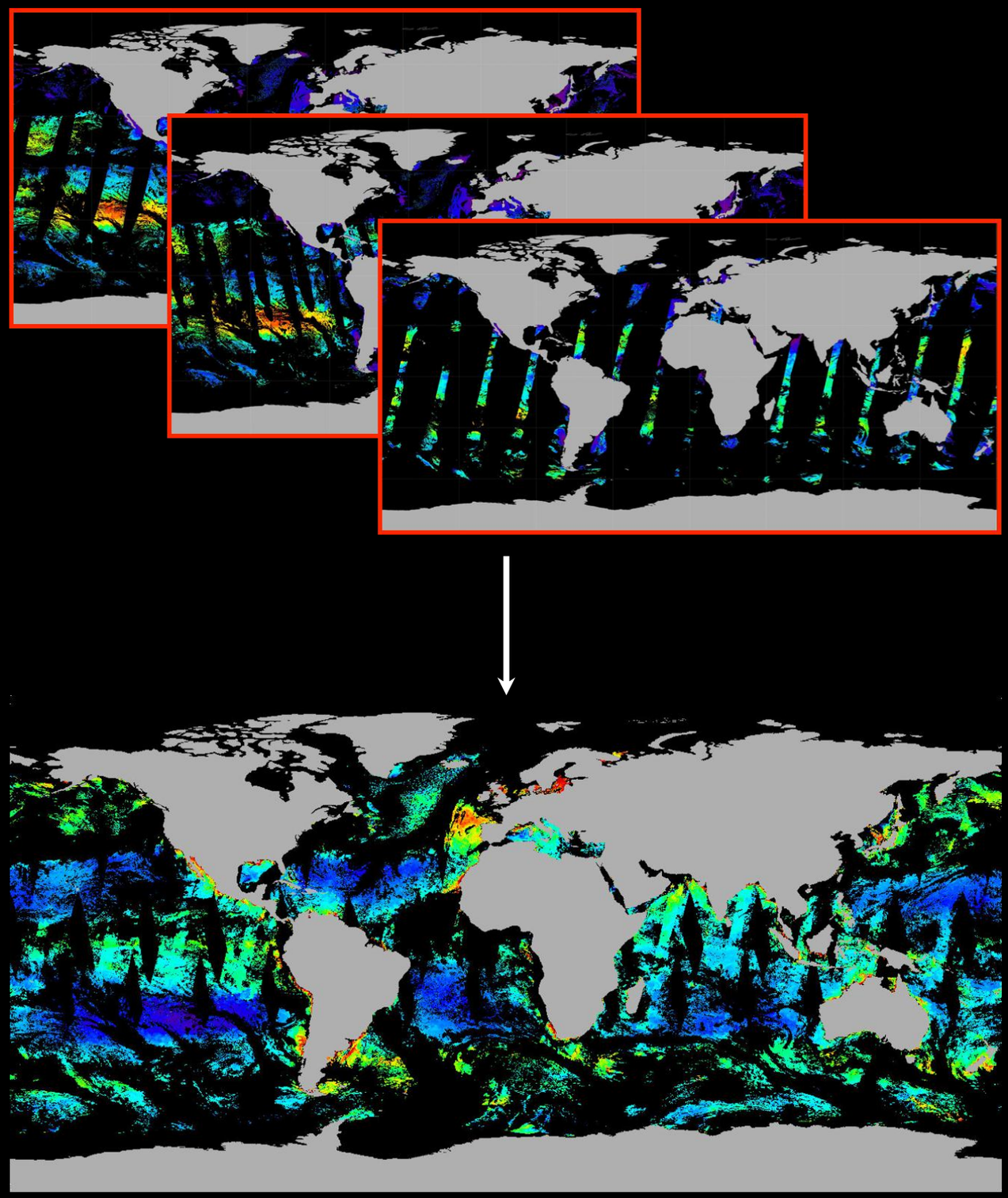
- filename patterns
  - if matches a `glob` expression
- version control
  - if == a committed version in a repository
- transitivity
  - if all antecedents are endorsed

manual endorsement

- endorse
  - optional comment
- ignore
  - object is irrelevant
- skip
  - punt for now

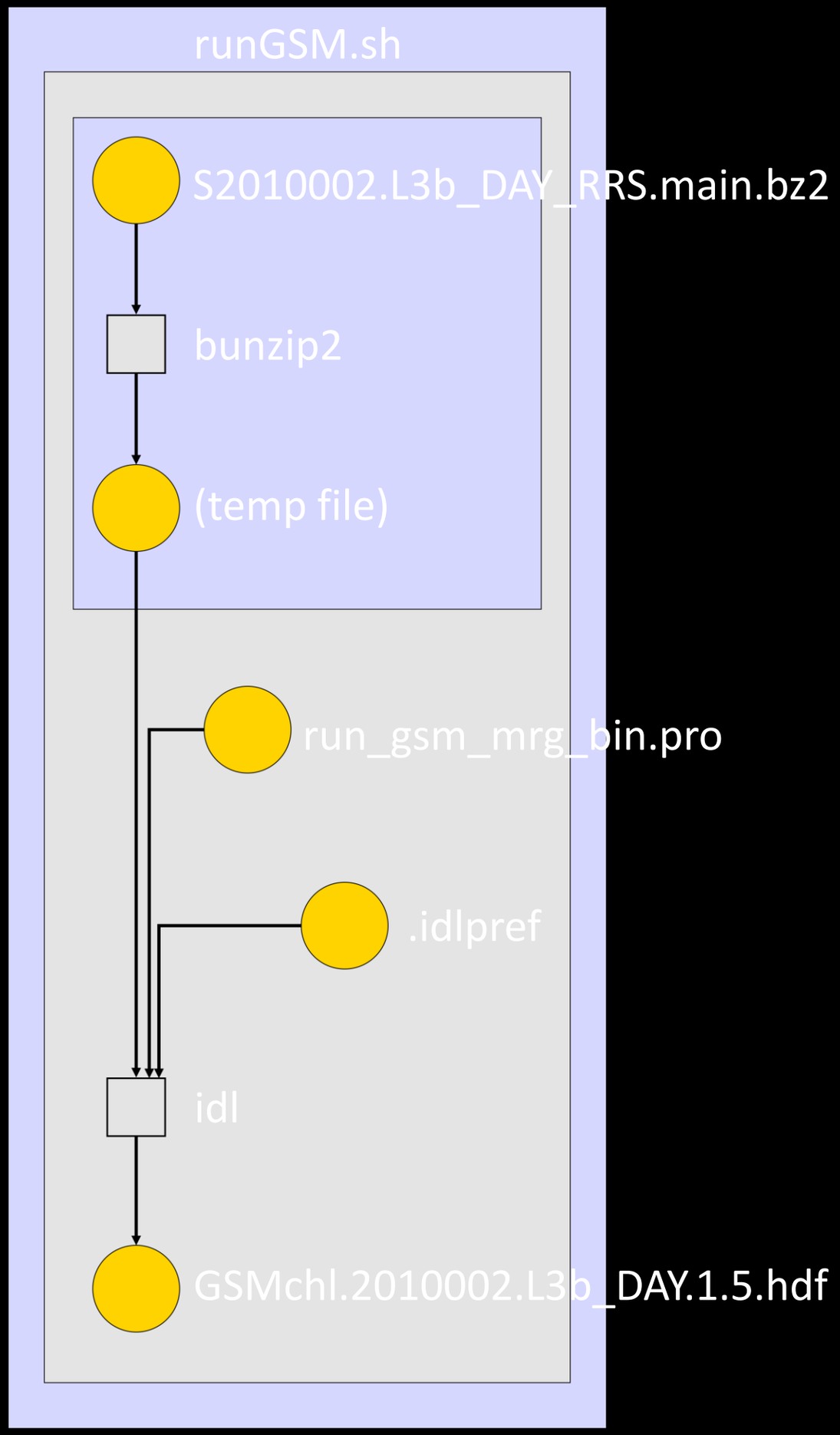
example

- ocean color algorithm



example

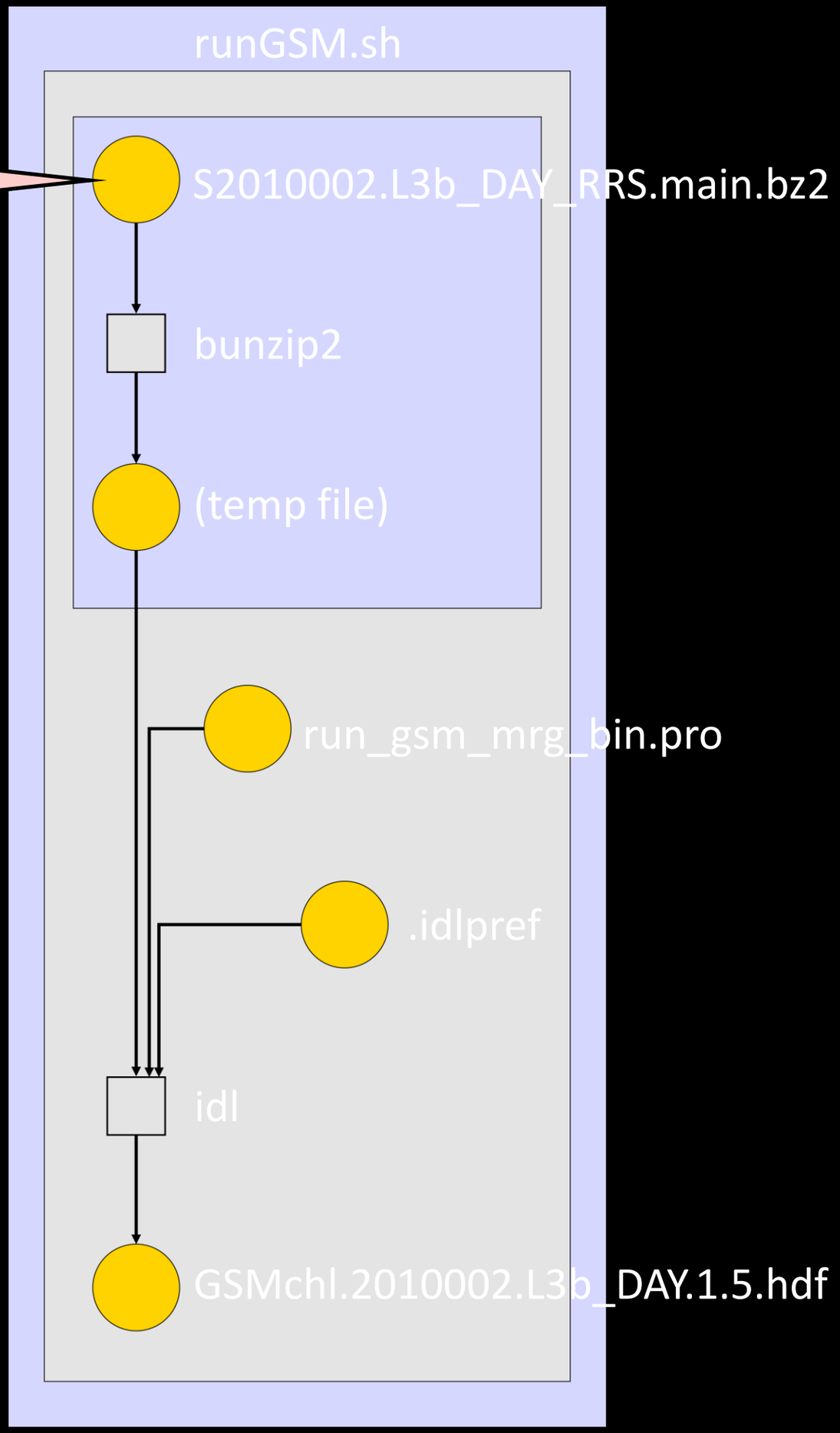
- ocean color algorithm
- provenance captured by ES3; rendered as dataflow graph
- now, let's *publish*→



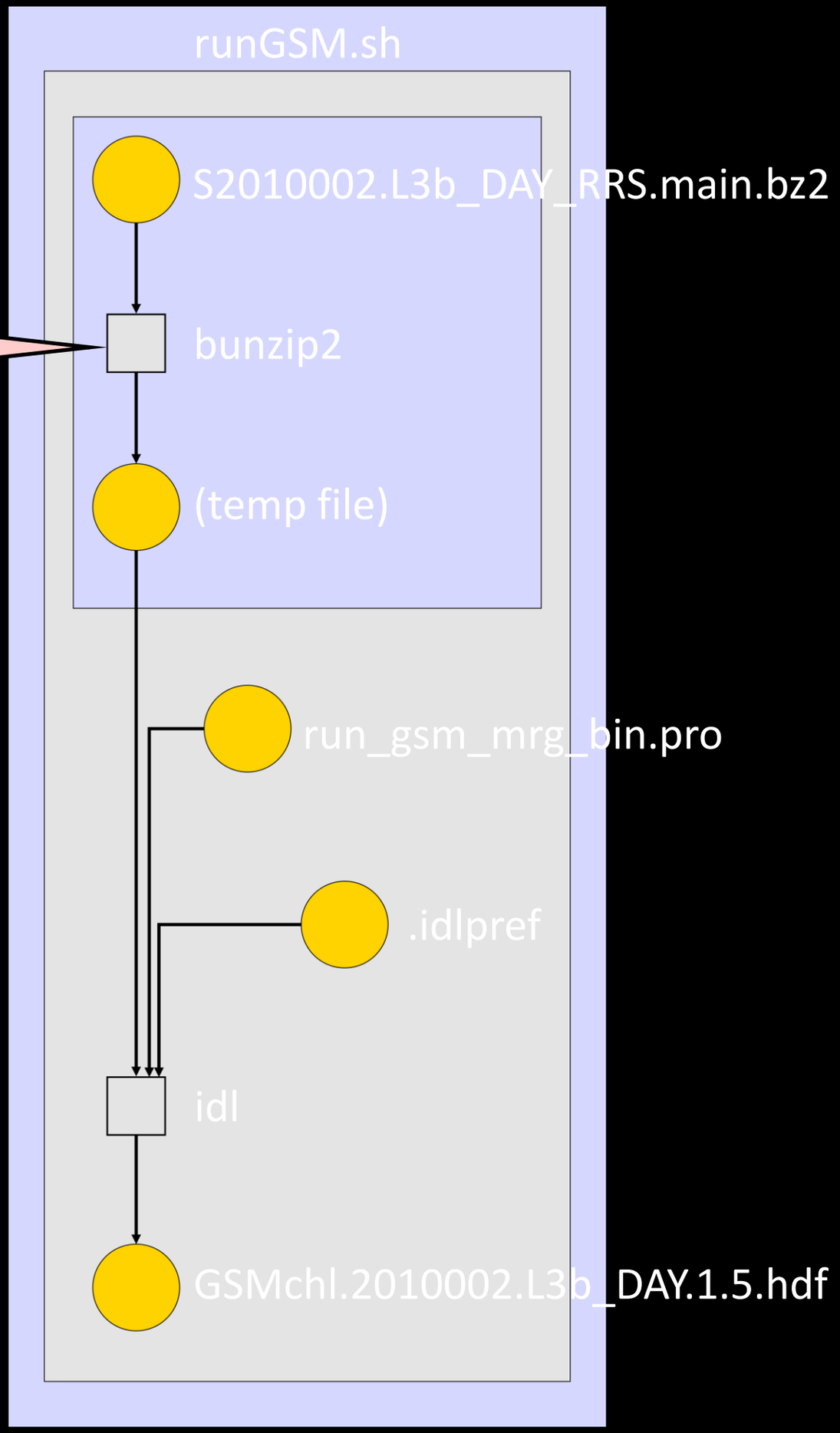
Endorsed by: glob rule: `*/data/*.bz2`

Annotation: SeaWiFS reprocessing 5.2

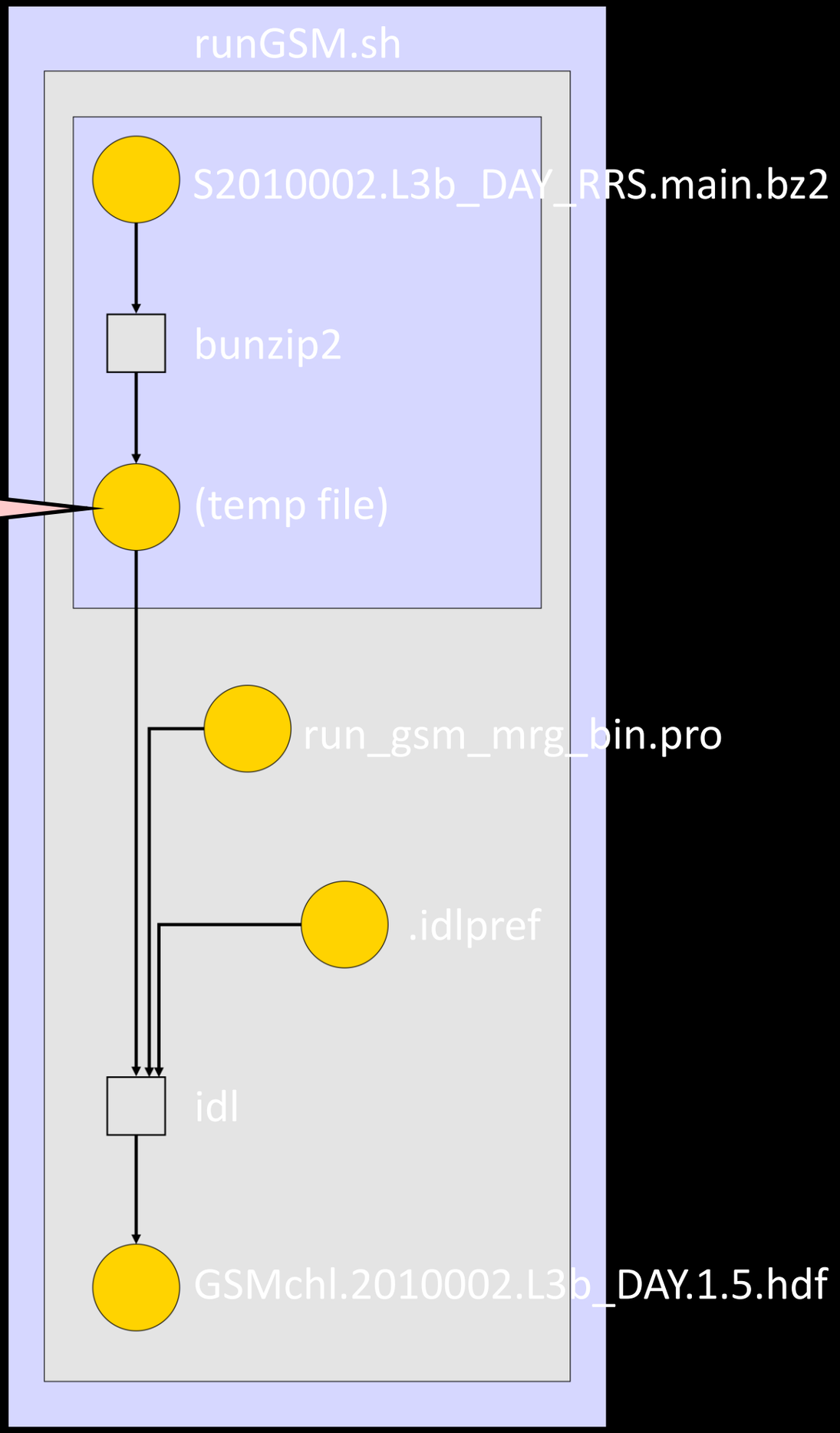
Assertion check: file (at time referenced) was unchanged since rule creation (failed assertion may indicate annotation is incorrect)



Endorsed by: glob rule: `/usr/bin/*`
Annotation: operating system tool
Assertion check: (none)



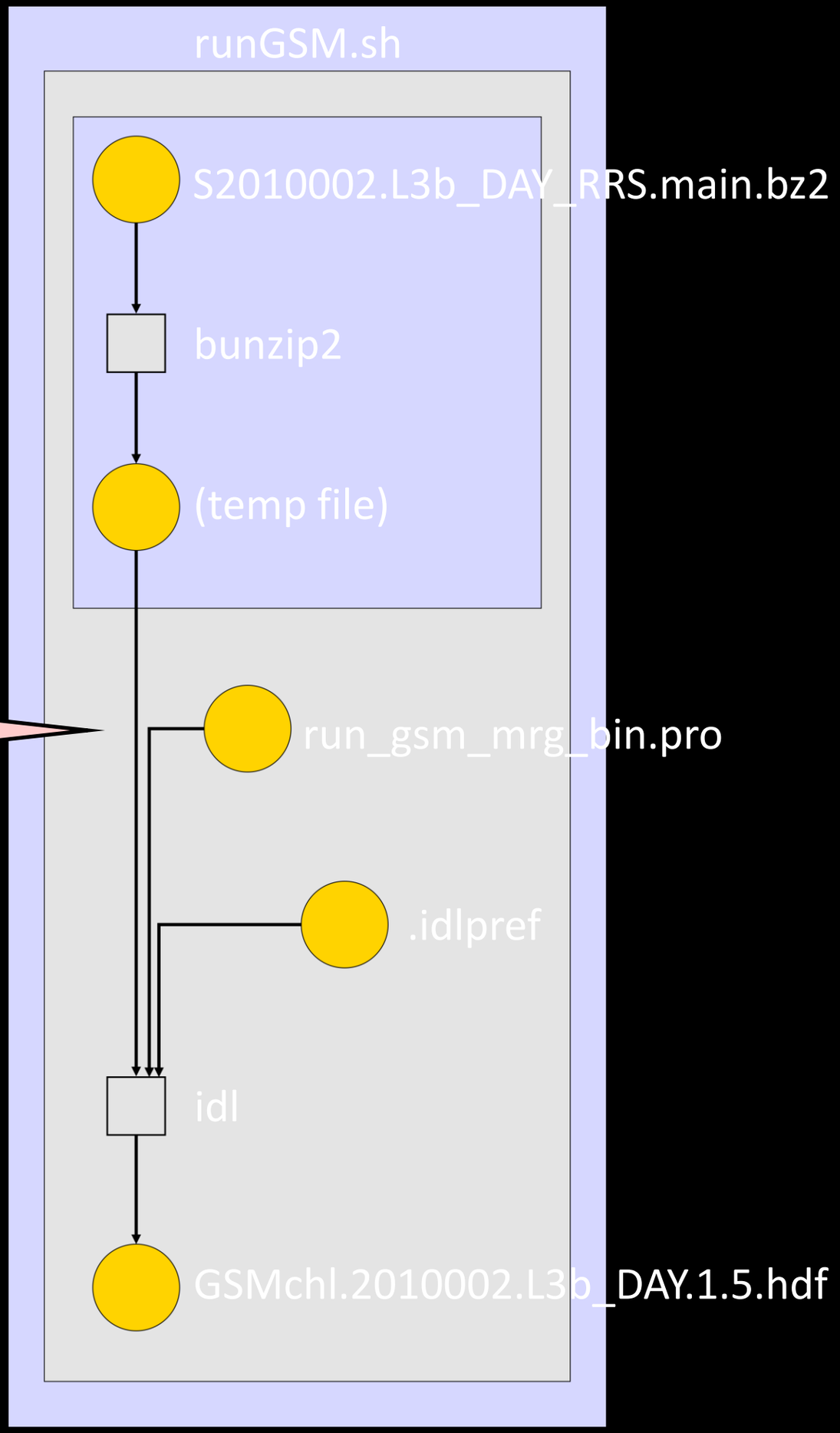
Endorsed by: transitivity

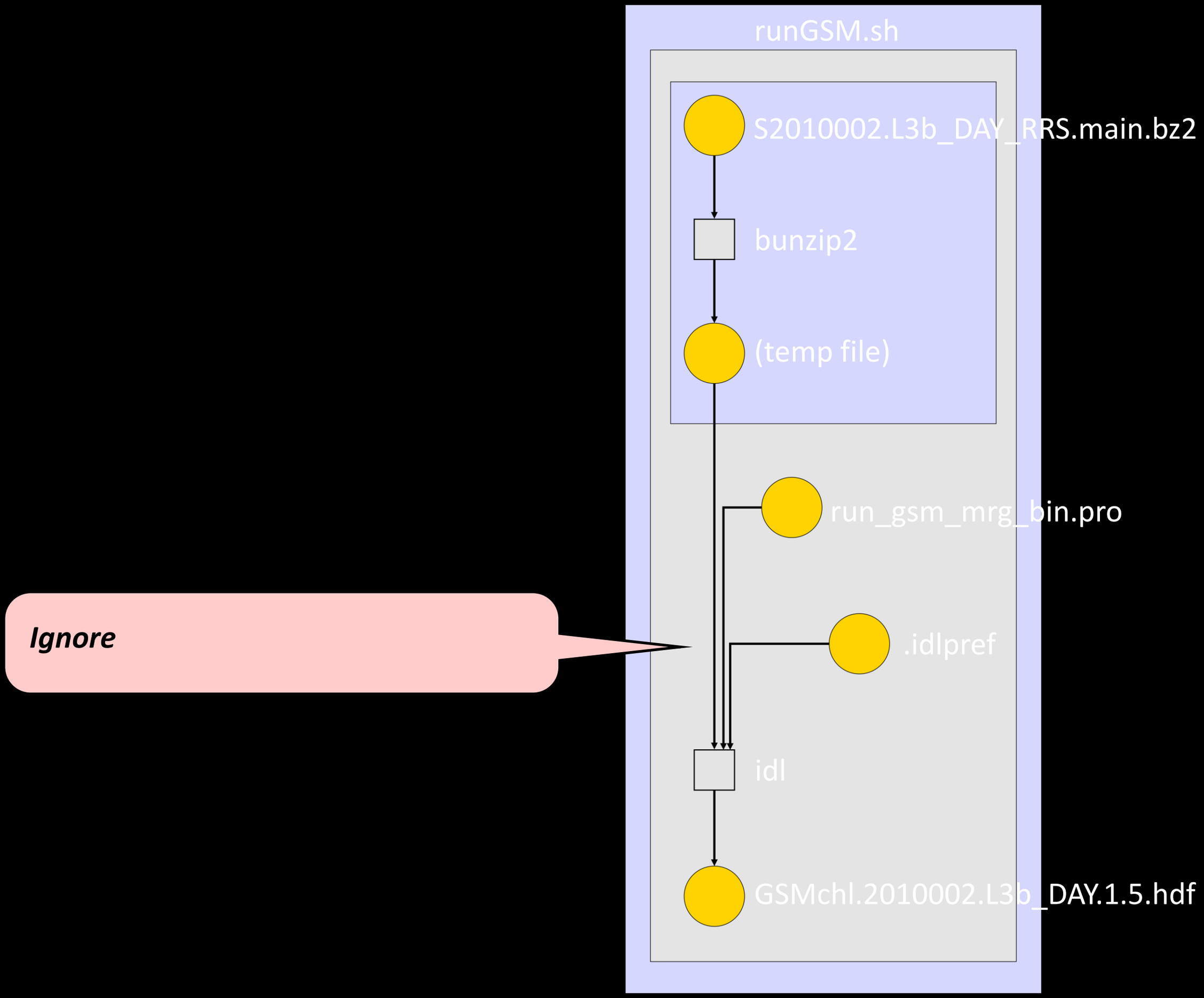


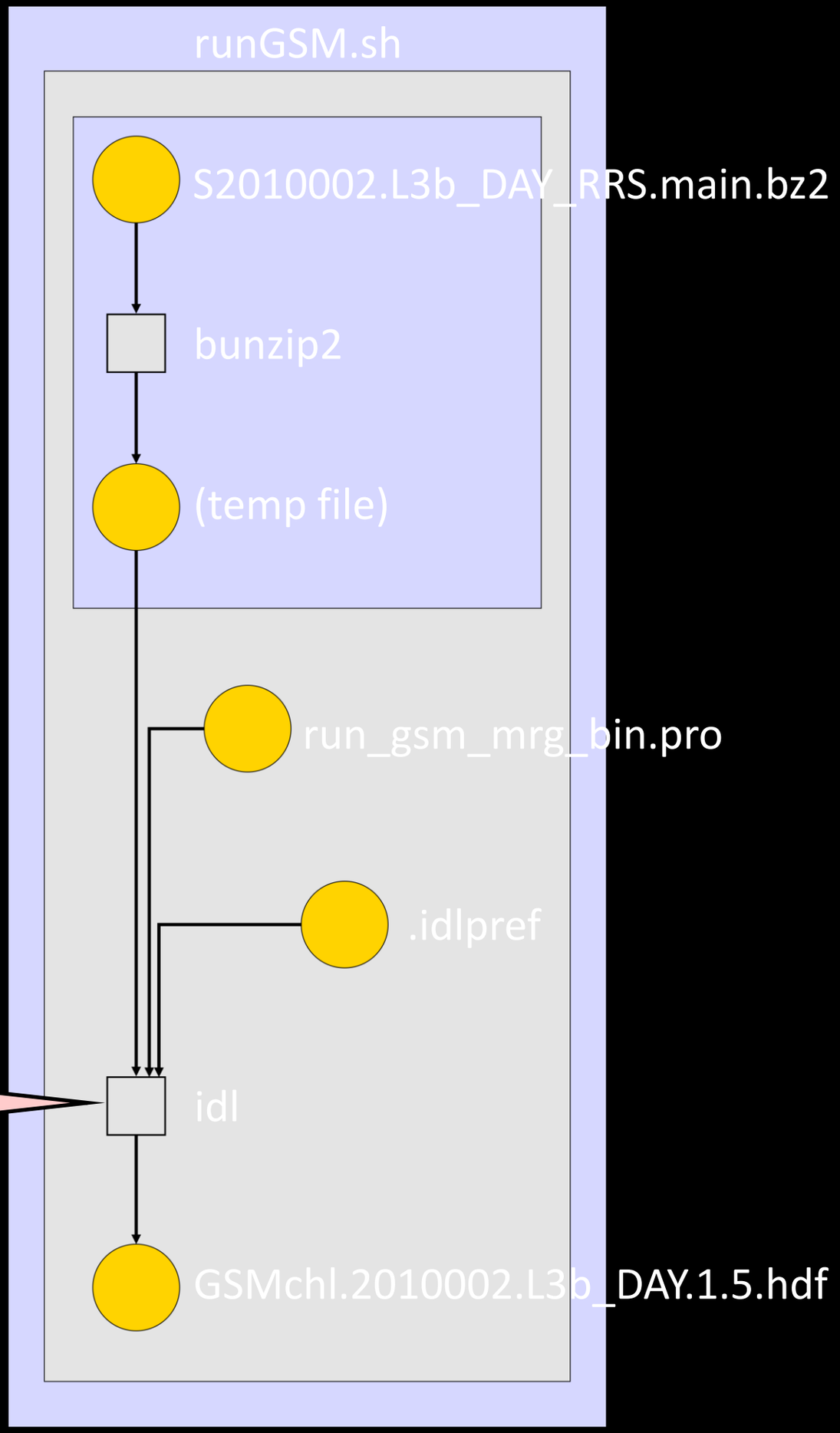
Endorsed by: version control system rule

Annotation: GSMS

Assertion check: file (at time referenced) corresponded to committed version (failed assertion may indicate uncommitted code was used)



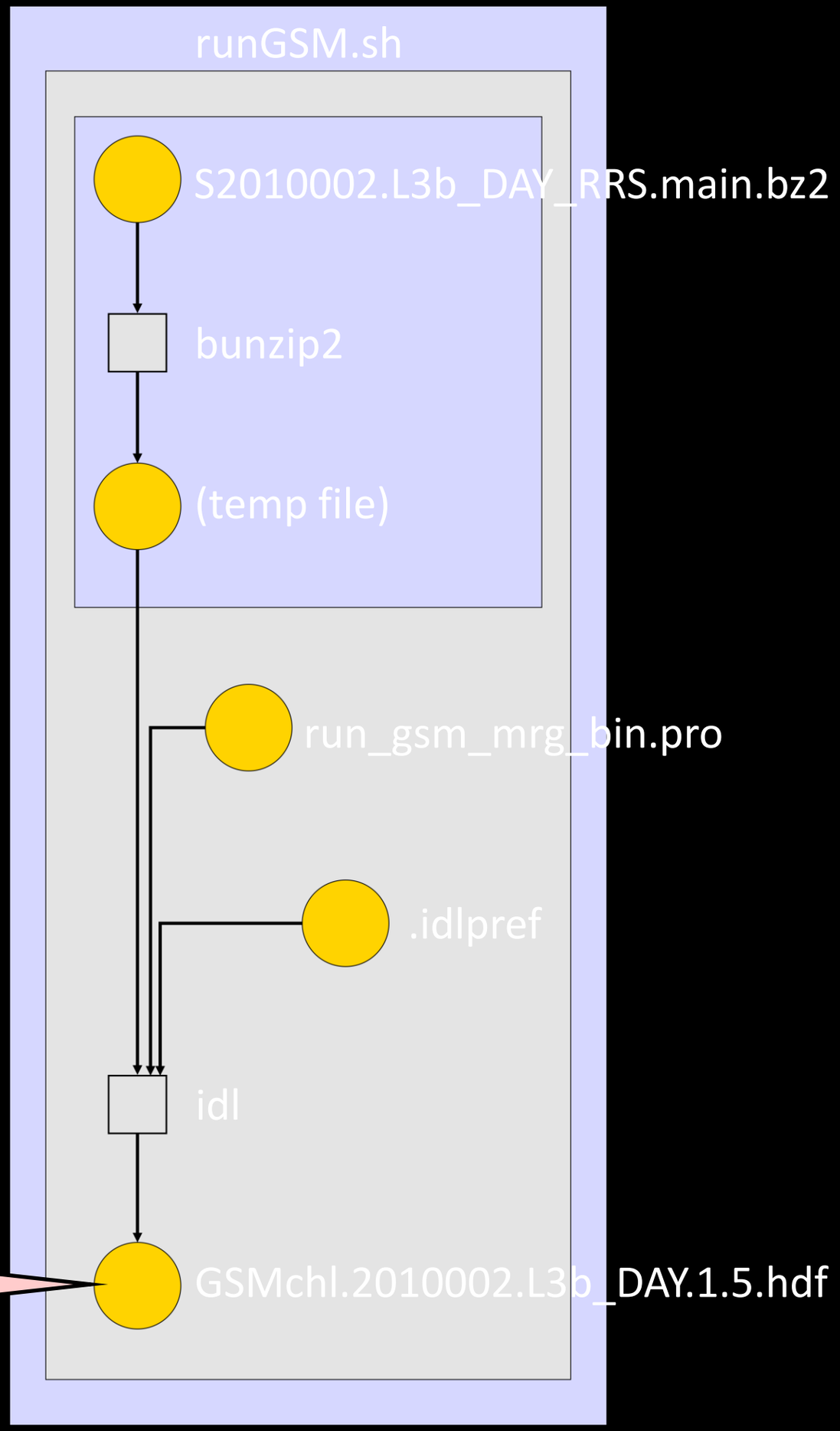

runGSM.sh
S2010002.L3b_DAY_RRS.main.bz2
bunzip2
(temp file)
run_gsm_mrg_bin.pro
.idlpref
idl
GSMchl.2010002.L3b_DAY.1.5.hdf
Ignore



Endorsed by: glob rule: `/itt/idl/*`

Annotation: IDL 8.0

Assertion check: file (at time referenced) was unchanged since rule creation (failed assertion may indicate annotation is incorrect)



Endorsed by: transitivity

issues & next steps

- granularity
  - read/write file → provenance graph cycle
- compilation
  - versioned source vs. executed binary
- distributed version control
  - single “version” in multiple changesets

Thanks!