Characterization in workflows

André Schaaff, CDS

Team

- Collegial work involving
  - François Bonnarel, Brice Gassmann and Cyril Pestel, CDS
  - Mireille Louys, LSIIT
  - Eric Slesak, Observatoire de Nice
  - 2 trainees Grégory Mantelet and Omar Benjelloun
- Discussions in the frame of VO France Workflow working group

Workflow use cases...

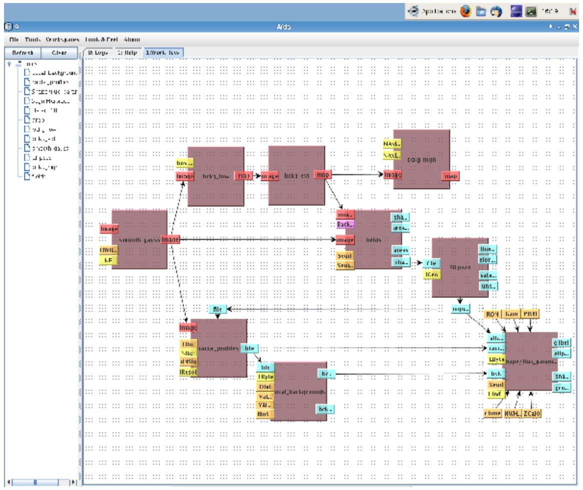
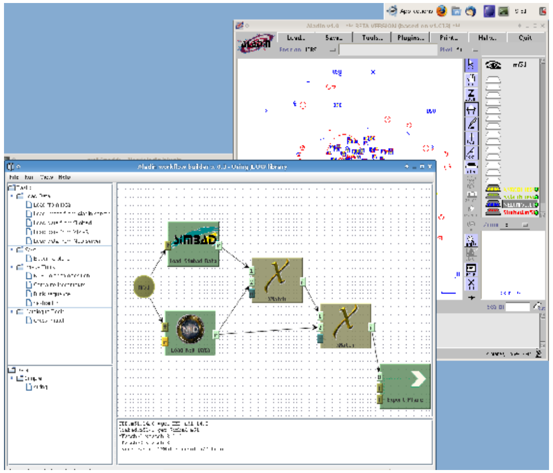
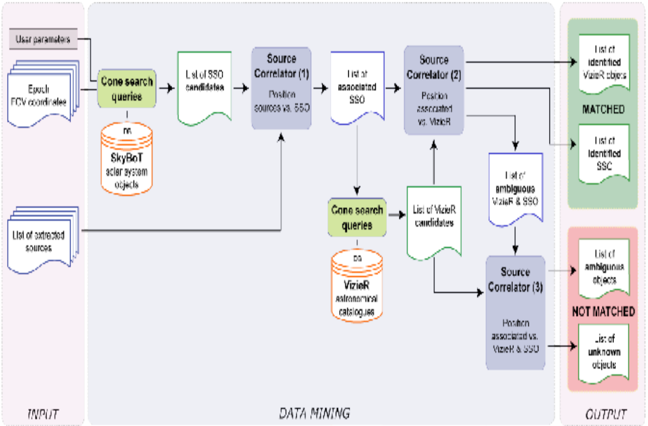


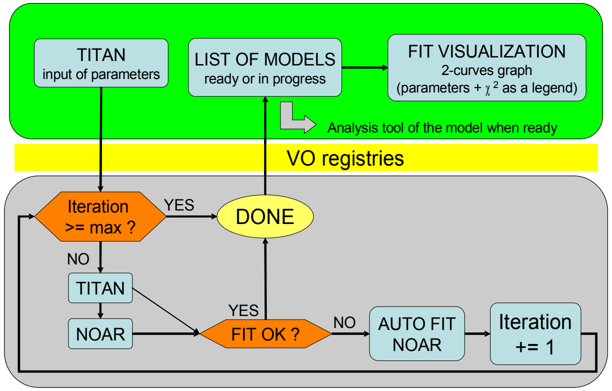
Image processing, E. Slezak.



Aladin scripting,C.Pestel, T. Boch.



Data Mining, J. Berthier et al.



TITAN/NOAR, L. Chevallier.

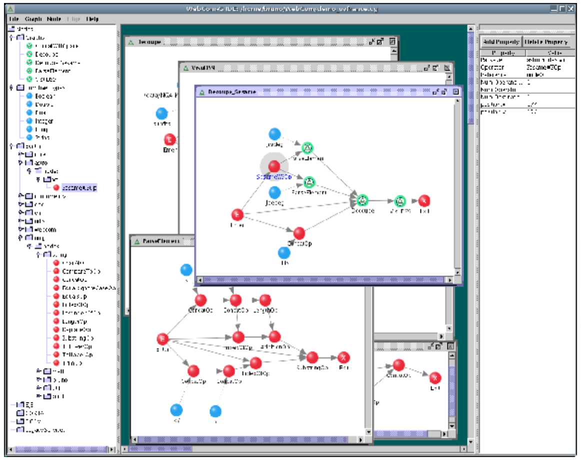
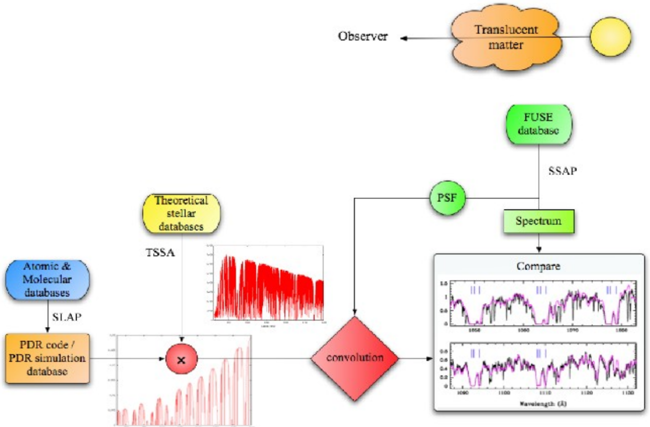


Image extraction from a catalogue, B. Voisin.

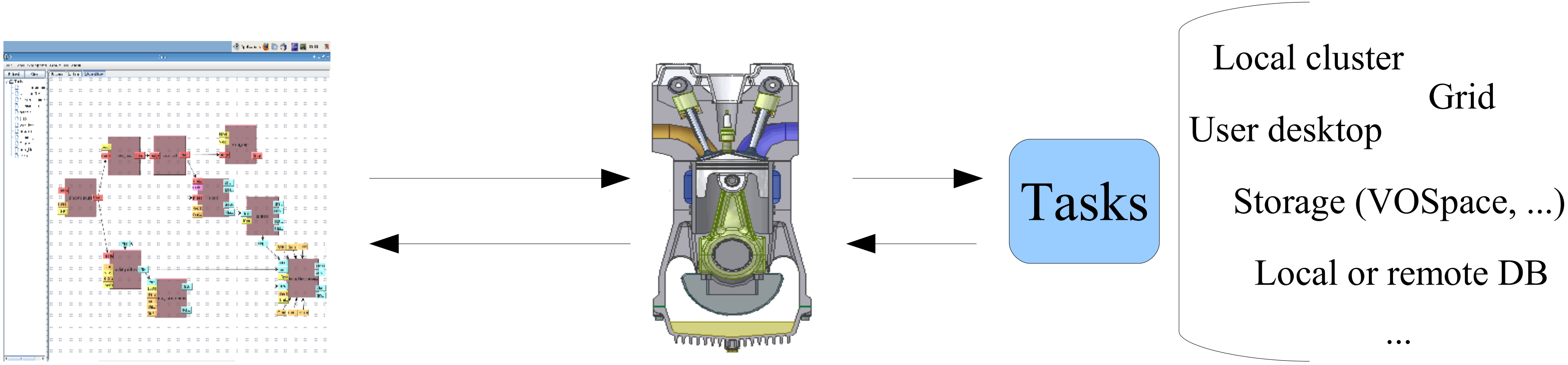


Simulation, F. Le Petit et al.

Characterization in workflows

Workflow systems

- “Sophisticated” workflow system
  - Graphical design tool
  - Workflow description (XML, ...) is sent to an engine who executes the workflow by dispatching the tasks
  - Execution is often visible step by step
  - Possible storage of intermediate data to change some parameters without the re execution of the whole workflow
  - Result(s) can be exploited through tools related to the kind of output data (FITS, ...)



Tasks

Local cluster

Grid

User desktop

Storage (VOSpace, ...)

Local or remote DB

...

Workflows in the VO

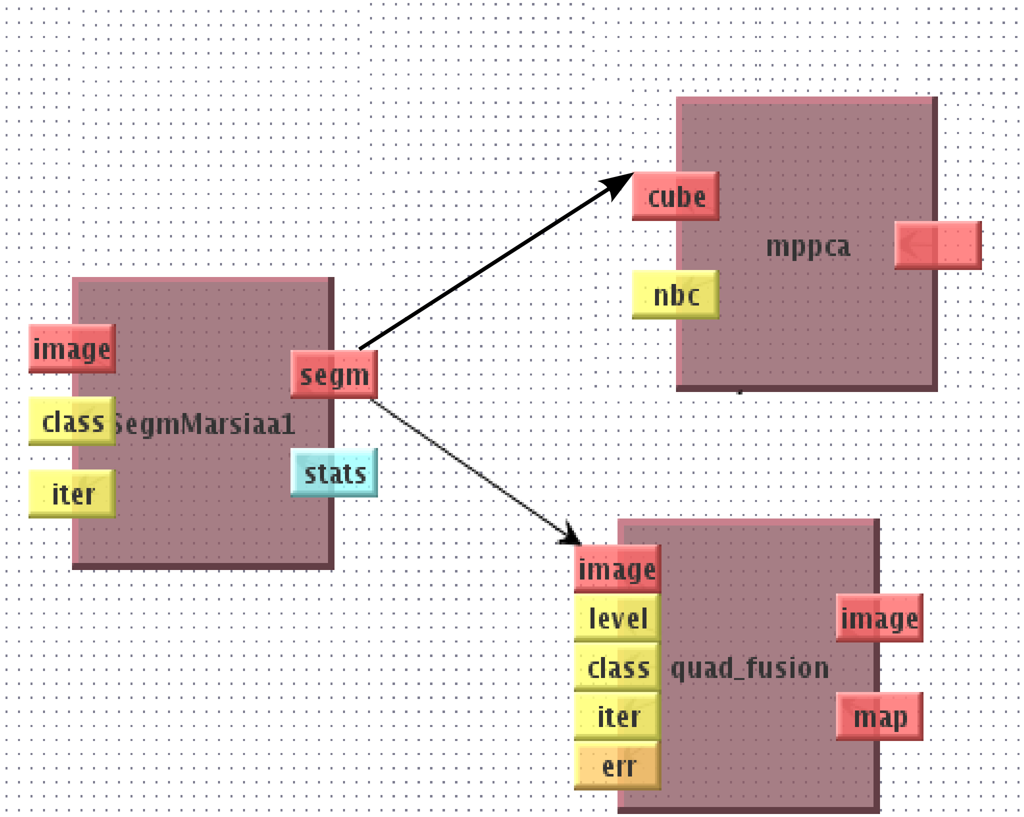
- Use and coordination of the services are possible through workflows
- Registry
  - Adaptive workflows with a choose of tools depending on parameters like the availability (see VOSI), ...
- VOSpace
  - Storage of intermediate (deleted after each execution or temporary conserved to replay partially the workflow, ...) or final data produced during the workflow execution, ...
- UWS
  - Use of asynchronous VO services in a workflow, ...
- ...

Common problems in workflows

- **Applications called in workflows are often developed by different persons, with different languages, on different systems, ...**
  - **No unified error management, job failure, etc.**
- **...**
- **A workflow can involve computing resources like clusters, grids, access to databases, ...**
  - **For a 9 steps workflow if the step 6 requires a few hours (or days) of computing and the step 7 crashes due (for example) to a bad entry value, the workflow will probably end...**
    - A workflow process is dependant from its composition
    - How to reduce this ? (investment in CPU, user time, ...)

How to reduce this ?

- **Checking of a workflow before and during its execution ?**



- **Benefits**
  - A part of the checking is done on the client side before the submission to the engine
  - Minimize the use of the external resources if validation fails
  - Optimization of the user time
  - ...

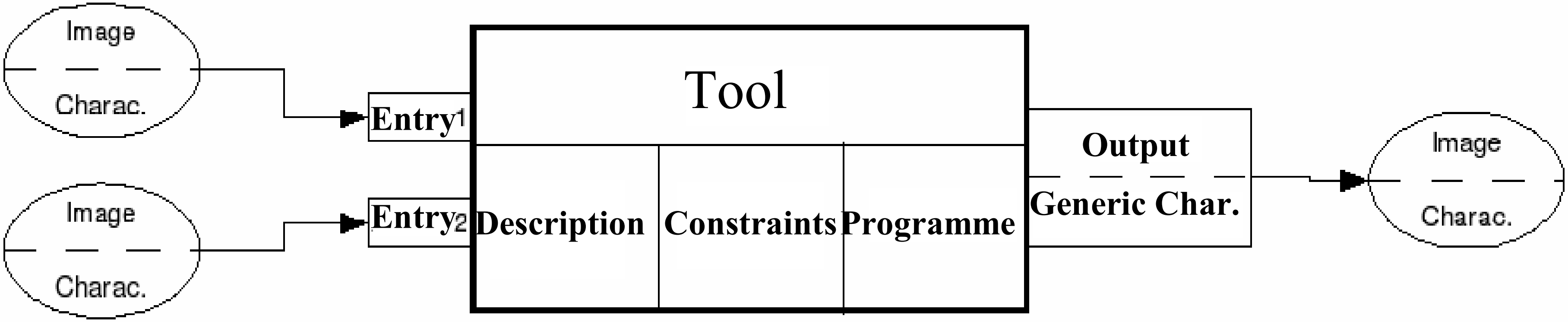
First step

- **Checking of the inputs/outputs**
  - **At a low level : verify the types of the linked I/O**
  - **Better : go further and check more than the type**
  - **Try to do it for tools with FITS files as entries and use the Characterization standard**
    - FITS file + its characterization file
    - A constraints file for each concerned tool
    - Add a characterization file/ contraints checker to the Workflow tool
  - **Do this checking also during to the execution**
    - Generate a characterization file for a FITS file resulting from the execution

IVOA Characterization

- **From the last reference document**
  - *This document defines the high level metadata necessary to describe the physical parameter space of observed or simulated astronomical data sets, such as 2D-images, data cubes, X-ray event lists, IFU data, etc...The Characterisation data model is an abstraction which can be used to derive a structured description of any relevant data and thus to facilitate its discovery and scientific interpretation. The model aims at facilitating the manipulation of heterogeneous data in any VO framework or portal.*

Recapitulation



- **Before the execution**
  - **Constraints on entries are defined for each tool**
  - **A validation step checks the entries**
- **During the execution**
  - **After the step i, a characterization file is generated for the outputs and checked with the step i+1 constraints before its execution**

Workflow test bed

AÏDA, Astronomical Image processIng Distribution Architecture

Contributors

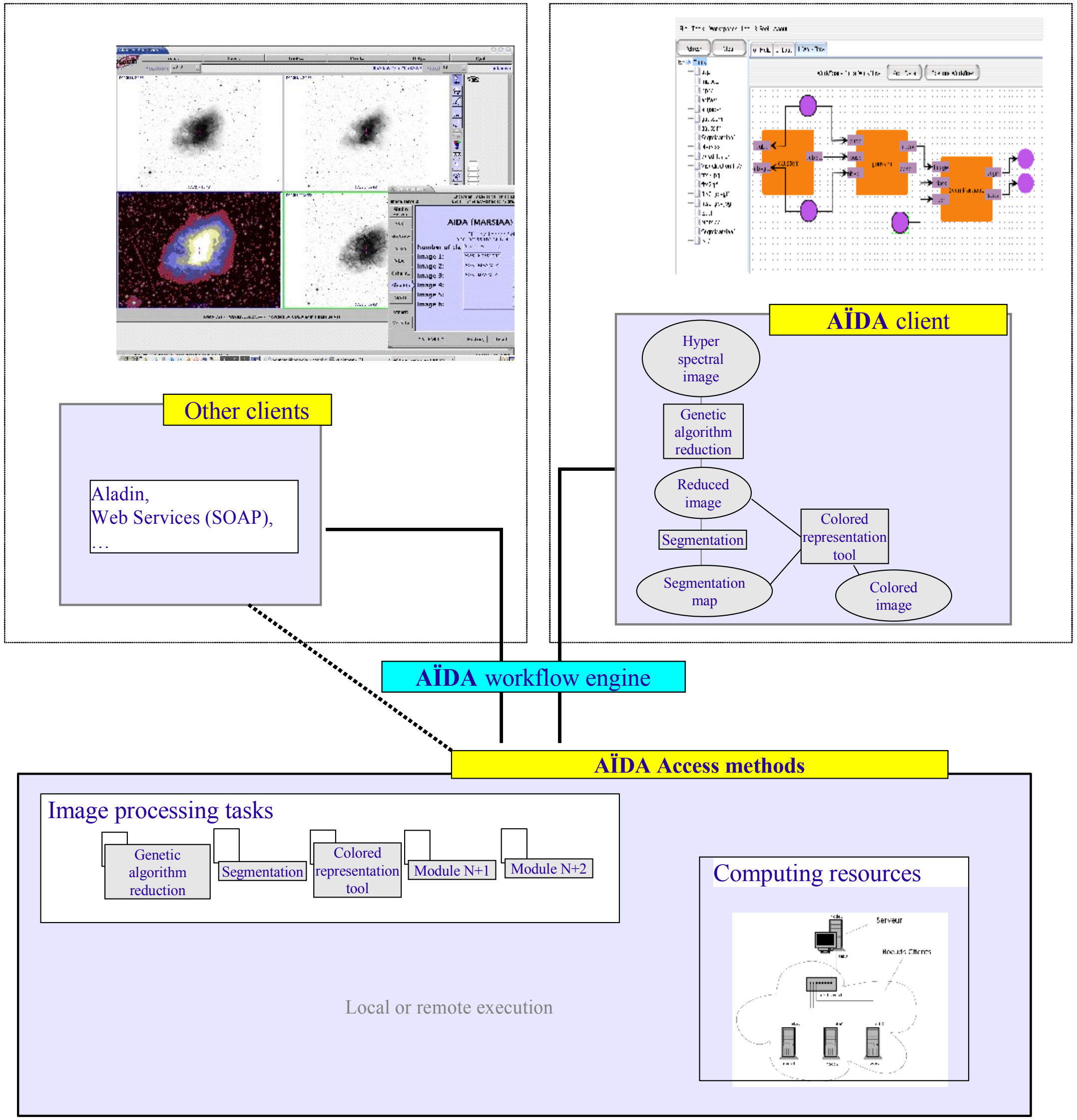
O. Benjelloun, characterization integration
J. Beugnot*, packaging
F. Bonnarel, architecture
J.-J. Claudon*, core development
B. Gassmann, characterization & Camea
M. Louys, architecture
G. Mantelet*, characterization integration
C. Pestel, JLOW - design capabilities, new developments
A. Schaaff, architecture

CDS & LSIIT

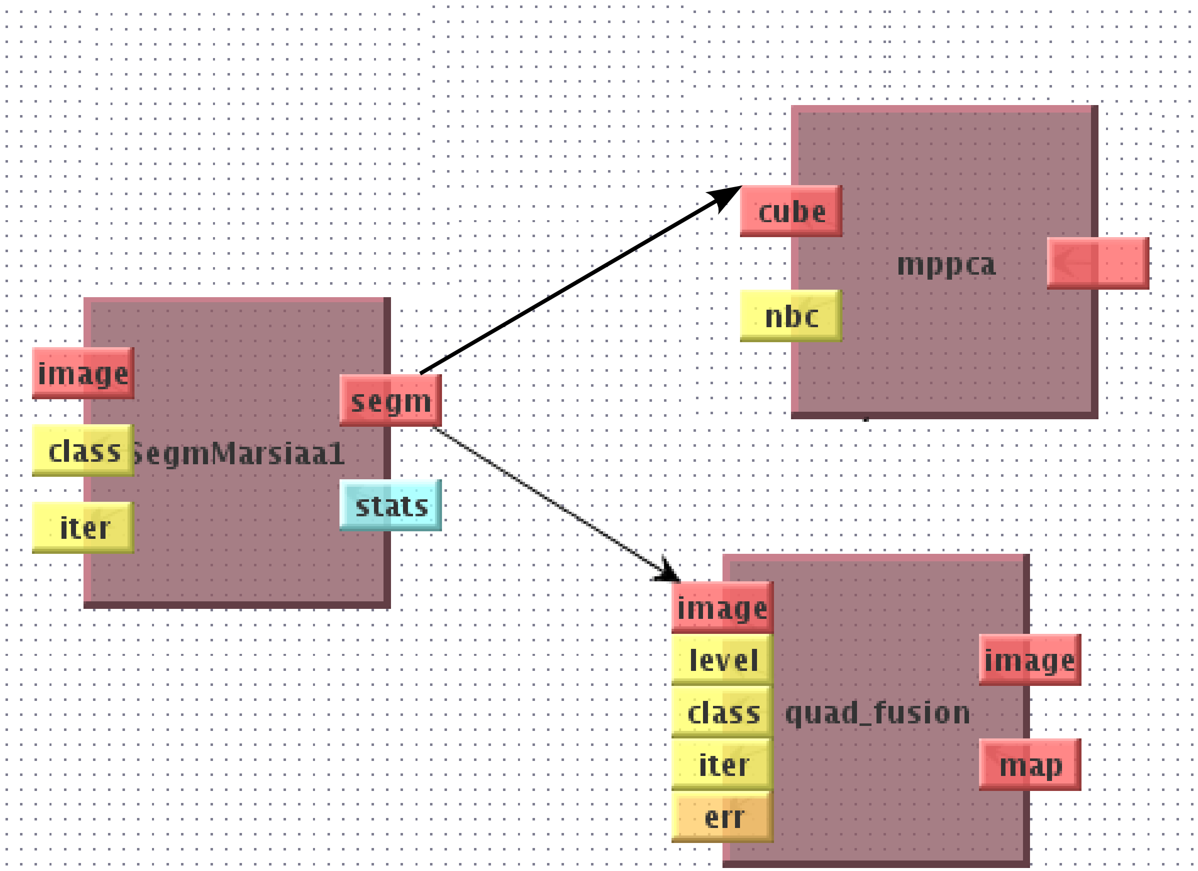
E. Slezak, Use cases
Observatoire de Nice

(have left)*

Work done in the frame of the French « **Massive Data in Astronomy** » project (2003-2006), **VO France** and **VOTECH**

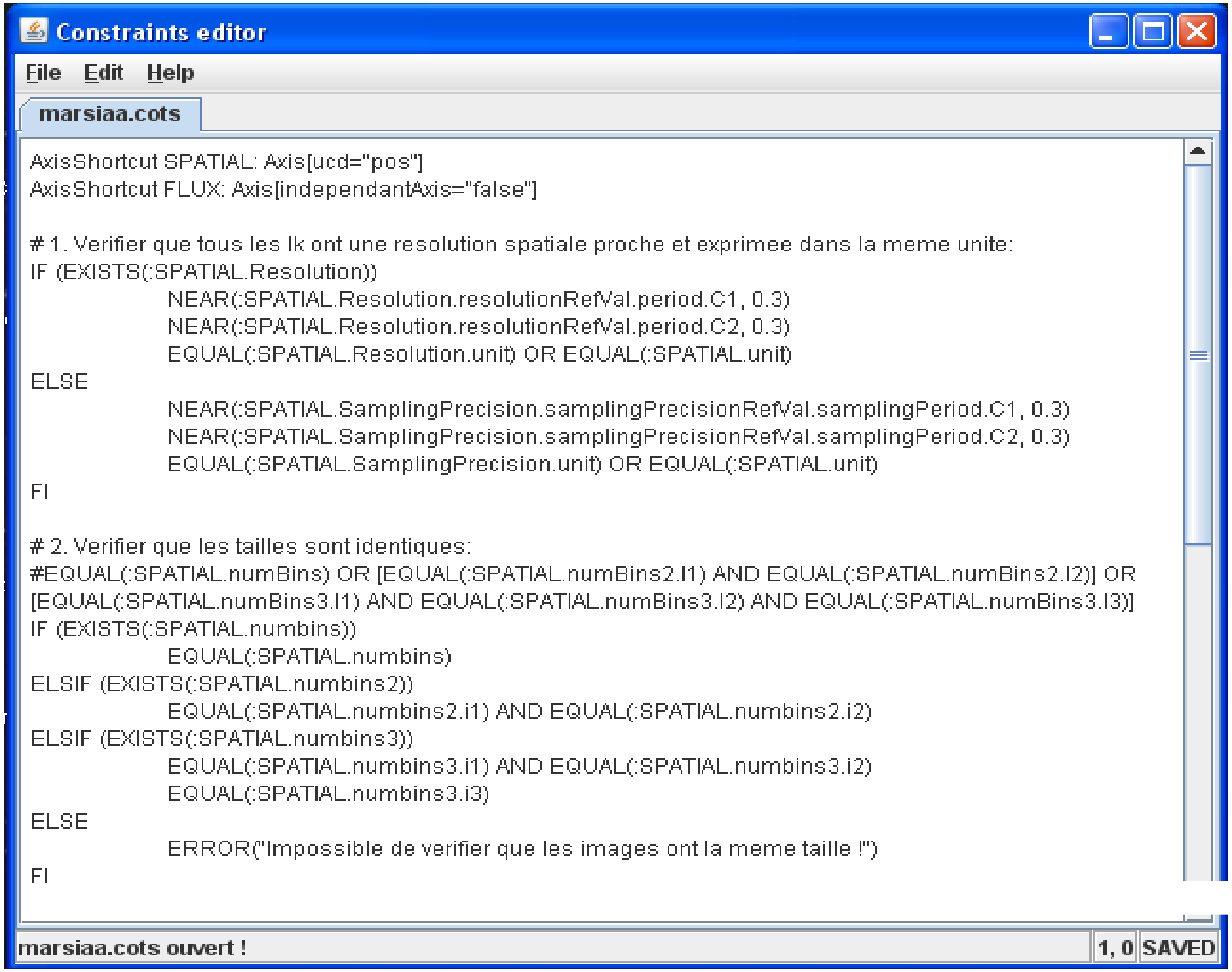


We need a use case



Second step : write the constraints for each tool

- **We have added a simple constraints editor to AÏDA ...**



```
AxisShortcut SPATIAL: Axis[ucd="pos"]
AxisShortcut FLUX: Axis[independantAxis="false"]

# 1. Verifier que tous les Ik ont une resolution spatiale proche et exprimee dans la meme unite:
IF (EXISTS(:SPATIAL.Resolution))
                NEAR(:SPATIAL.Resolution.resolutionRefVal.period.C1, 0.3)
                NEAR(:SPATIAL.Resolution.resolutionRefVal.period.C2, 0.3)
                EQUAL(:SPATIAL.Resolution.unit) OR EQUAL(:SPATIAL.unit)
ELSE
                NEAR(:SPATIAL.SamplingPrecision.samplingPrecisionRefVal.samplingPeriod.C1, 0.3)
                NEAR(:SPATIAL.SamplingPrecision.samplingPrecisionRefVal.samplingPeriod.C2, 0.3)
                EQUAL(:SPATIAL.SamplingPrecision.unit) OR EQUAL(:SPATIAL.unit)
FI

# 2. Verifier que les tailles sont identiques:
#EQUAL(:SPATIAL.numBins) OR [EQUAL(:SPATIAL.numBins2.I1) AND EQUAL(:SPATIAL.numBins2.I2)] OR
[EQUAL(:SPATIAL.numBins3.I1) AND EQUAL(:SPATIAL.numBins3.I2) AND EQUAL(:SPATIAL.numBins3.I3)]
IF (EXISTS(:SPATIAL.numbins))
                EQUAL(:SPATIAL.numbins)
ELSIF (EXISTS(:SPATIAL.numbins2))
                EQUAL(:SPATIAL.numbins2.i1) AND EQUAL(:SPATIAL.numbins2.i2)
ELSIF (EXISTS(:SPATIAL.numbins3))
                EQUAL(:SPATIAL.numbins3.i1) AND EQUAL(:SPATIAL.numbins3.i2)
                EQUAL(:SPATIAL.numbins3.i3)
ELSE
                ERROR("Impossible de verifier que les images ont la meme taille !")
FI
```

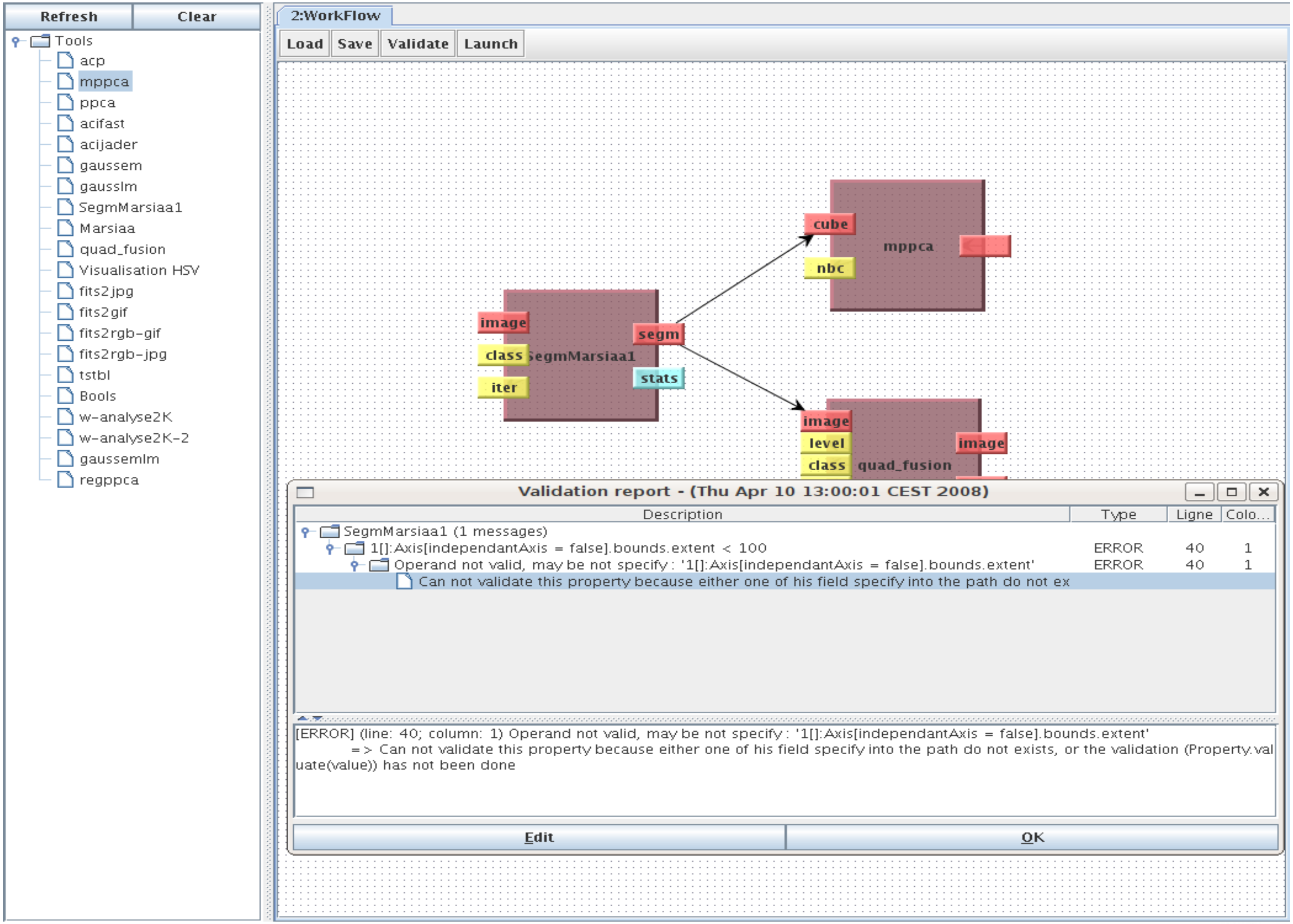
marsiaa.cots ouvert !   1, 0 SAVED

Definition of the constraints

- ... and defined the grammar to generate the constraints parser
  - Very close to our needs (and to Characterization)

```
 AxisShortcut SPATIAL: Axis[ucd="pos"]
AxisShortcut FLUX: Axis[independantAxis="false"]

# 1. Verify that all the Ik have a close spatial resolution and are expressed in the same unit
IF (EXISTS(:SPATIAL.Resolution))
   NEAR(:SPATIAL.Resolution.resolutionRefVal.period.C1, 0.3)
   NEAR(:SPATIAL.Resolution.resolutionRefVal.period.C2, 0.3)
   EQUAL(:SPATIAL.Resolution.unit) OR EQUAL(:SPATIAL.unit)
ELSE
   NEAR(:SPATIAL.SamplingPrecision.samplingPrecisionRefVal.samplingPeriod.C1, 0.3)
   NEAR(:SPATIAL.SamplingPrecision.samplingPrecisionRefVal.samplingPeriod.C2, 0.3)
   EQUAL(:SPATIAL.SamplingPrecision.unit) OR EQUAL(:SPATIAL.unit)
FI

# 2. Verify if the sizes are identical
IF (EXISTS(:SPATIAL))
   EQUAL(:SPATIAL.numbins)
ELSIF (EXISTS(:SPATIAL.numbins2))
   EQUAL(:SPATIAL.numbins2.i1) AND EQUAL(:SPATIAL.numbins2.i2)
ELSIF (EXISTS(:SPATIAL.numbins3))
   EQUAL(:SPATIAL.numbins3.i1) AND EQUAL(:SPATIAL.numbins3.i2)
   EQUAL(:SPATIAL.numbins3.i3)
ELSE
   ERROR("Impossible de vérifier que les images ont la même taille !")
FI

# 3. Vérifier que toutes les images sont superposables
EQUAL(:SPATIAL.Coverage.location.unit) OR EQUAL(:SPATIAL.Coverage.unit) OR
EQUAL(:SPATIAL.unit)
EQUAL(:SPATIAL.Coverage.location.coord_system_id)

# 4. Observable : (min-max)>=100 else WARNING
EQUAL(1[]:FLUX.coverage.bounds.unit) OR EQUAL(:SPATIAL.Coverage.unit) OR
EQUAL(:SPATIAL.unit)
IF (1[]:FLUX.bounds.limitHi - 1[]:FLUX.bounds.limitLo >= 100)
   WARNING("(Observables: min-max <100) Il faut faire une normalisation en niveau de gris !")
FI

# 5. ...
EQUAL(:FLUX.ucd)
1[]:FLUX.bounds.extent < 100
.....


 FI
```

Third step : validation report generation

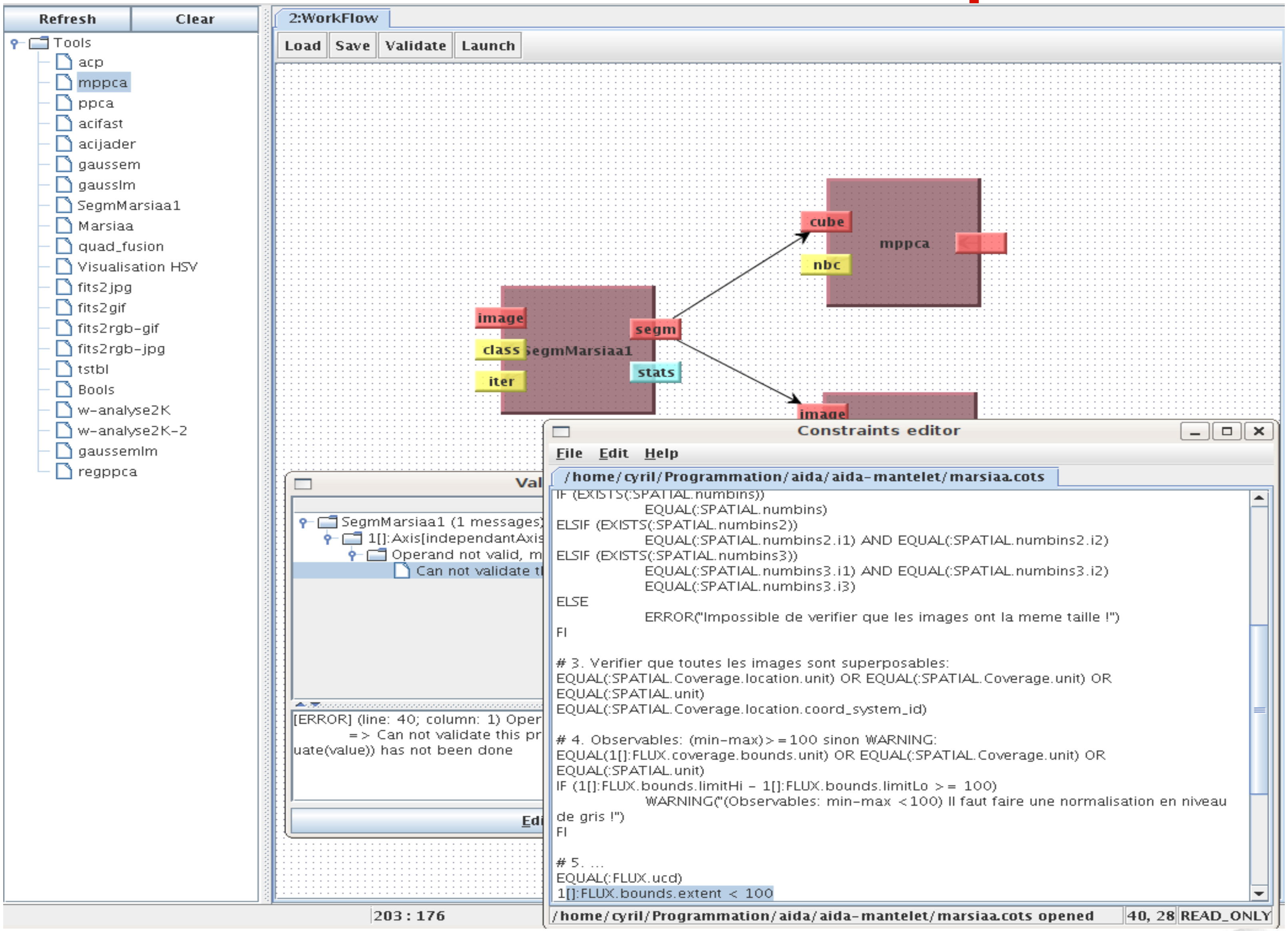


Validation report - (Thu Apr 10 13:00:01 CEST 2008)

Description	Type	Ligne	Colo...
SegmMarsiaa1 (1 messages)			
1[]:Axis[independantAxis = false].bounds.extent < 100	ERROR	40	1
Operand not valid, may be not specify : '1[]:Axis[independantAxis = false].bounds.extent'	ERROR	40	1
Can not validate this property because either one of his field specify into the path do not ex			

```
[ERROR] (line: 40; column: 1) Operand not valid, may be not specify : '1[]:Axis[independantAxis = false].bounds.extent'
        => Can not validate this property because either one of his field specify into the path do not exists, or the validation (Property.val
uate(value)) has not been done
```

AÏDA client with validation capabilities



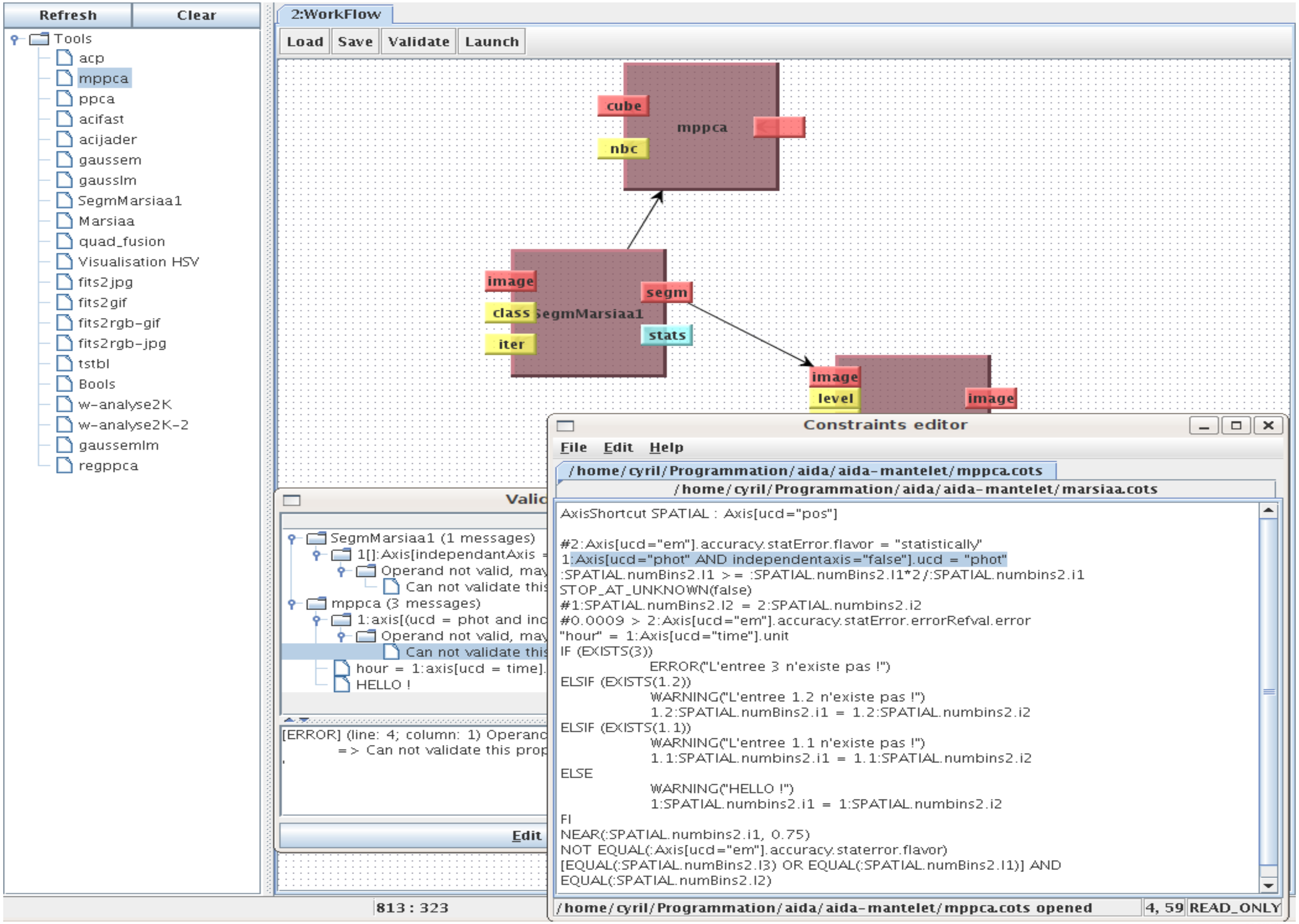
```
SegmMarsiaa1 (1 messages)
  1[]:Axis[independantAxis
    Operand not valid, m
      Can not validate t
```

```
[ERROR] (line: 40; column: 1) Oper
        => Can not validate this pr
uate(value)) has not been done
```

```
IF (EXISTS(:SPATIAL.numbins))
                EQUAL(:SPATIAL.numbins)
ELSIF (EXISTS(:SPATIAL.numbins2))
                EQUAL(:SPATIAL.numbins2.i1) AND EQUAL(:SPATIAL.numbins2.i2)
ELSIF (EXISTS(:SPATIAL.numbins3))
                EQUAL(:SPATIAL.numbins3.i1) AND EQUAL(:SPATIAL.numbins3.i2)
                EQUAL(:SPATIAL.numbins3.i3)
ELSE
                ERROR("Impossible de verifier que les images ont la meme taille !")
FI

# 3. Verifier que toutes les images sont superposables:
EQUAL(:SPATIAL.Coverage.location.unit) OR EQUAL(:SPATIAL.Coverage.unit) OR
EQUAL(:SPATIAL.unit)
EQUAL(:SPATIAL.Coverage.location.coord_system_id)

# 4. Observables: (min-max)>=100 sinon WARNING:
EQUAL(1[]:FLUX.coverage.bounds.unit) OR EQUAL(:SPATIAL.Coverage.unit) OR
EQUAL(:SPATIAL.unit)
IF (1[]:FLUX.bounds.limitHi - 1[]:FLUX.bounds.limitLo >= 100)
                WARNING("(Observables: min-max <100) Il faut faire une normalisation en niveau
de gris !")
FI

# 5. ...
EQUAL(:FLUX.ucd)
1[]:FLUX.bounds.extent < 100
```

/home/cyril/Programmation/aida/aida-mantelet/marsiaa.cots opened

AÏDA client with validation capabilities (2)



```
SegmMarsiaa1 (1 messages)
    1[]:Axis[independantAxis =
        Operand not valid, may
            Can not validate this
mppca (3 messages)
    1:axis[(ucd = phot and inc
        Operand not valid, may
            Can not validate this
    hour = 1:axis[ucd = time].
    HELLO !
```

```
[ERROR] (line: 4; column: 1) Operanc
        => Can not validate this prop
```

```
AxisShortcut SPATIAL : Axis[ucd="pos"]

#2:Axis[ucd="em"].accuracy.statError.flavor = "statistically"
1:Axis[ucd="phot" AND independentaxis="false"].ucd = "phot"
:SPATIAL.numBins2.I1 >= :SPATIAL.numBins2.I1*2/:SPATIAL.numbins2.i1
STOP_AT_UNKNOWN(false)
#1:SPATIAL.numBins2.I2 = 2:SPATIAL.numbins2.i2
#0.0009 > 2:Axis[ucd="em"].accuracy.statError.errorRefval.error
"hour" = 1:Axis[ucd="time"].unit
IF (EXISTS(3))
        ERROR("L'entree 3 n'existe pas !")
ELSIF (EXISTS(1.2))
        WARNING("L'entree 1.2 n'existe pas !")
        1.2:SPATIAL.numBins2.i1 = 1.2:SPATIAL.numbins2.i2
ELSIF (EXISTS(1. 1))
        WARNING("L'entree 1.1 n'existe pas !")
        1. 1:SPATIAL.numbins2.i1 = 1. 1:SPATIAL.numbins2.i2
ELSE
        WARNING("HELLO !")
        1:SPATIAL.numbins2.i1 = 1:SPATIAL.numbins2.i2
FI
NEAR(:SPATIAL.numbins2.i1, 0.75)
NOT EQUAL(:Axis[ucd="em"].accuracy.staterror.flavor)
[EQUAL(:SPATIAL.numBins2.I3) OR EQUAL(:SPATIAL.numBins2.I1)] AND
EQUAL(:SPATIAL.numBins2.I2)
```

/home/cyril/Programmation/aida/aida-mantelet/mppca.cots opened | 4, 59 | READ_ONLY

Ongoing work

- **Characterization generation from FITS files, example : 003.7858-39.2202.fits + <u>MappingSpecificAxis.map</u> ----> 003.7858-39.2202.uty**

```
AXIS1NAM + SpatialAxis.AxisName
AXIS1UCD + SpatialAxis.ucd
AXIS1UNI + SpatialAxis.unit
AXIS1CAL + SpatialAxis.calibrationStatus
AXIS1SYS + SpatialAxis.coordsystem
AXIS1STE + SpatialAxis.accuracy.statError.ErrorRefval.ErrorRefValue
AXIS1SYE + SpatialAxis.accuracy.sysError.ErrorRefval.ErrorRefValue
AXIS1IND + SpatialAxis.independentaxis
AXIS1BIN + SpatialAxis.numBins
AXIS1UND + SpatialAxis.undersamplingStatus
AXIS1REG + SpatialAxis.regularsamplingStatus
POSITIO1 + SpatialAxis.coverage.location.coord.Position2D.Value2.C1
POSITIO2 + SpatialAxis.coverage.location.coord.Position2D.Value2.C2
LOWERBOX + SpatialAxis.coverage.bounds.limits.Coord2VecInterval.LoLimit2Vec
UPPERBOX + SpatialAxis.coverage.bounds.limits.Coord2VecInterval.HiLimit2Vec
SEEING   + SpatialAxis.resolution.resolutionRefVal
PIXSCALE + SpatialAxis.samplingPrecision.samplingPrecisionRefVal.samlingPeriod
AXIS2NAM + TimeAxis.AxisName
AXIS2UCD + TimeAxis.ucd
AXIS2UNI + TimeAxis.unit
AXIS2CAL + TimeAxis.calibrationStatus
AXIS2SYS + TimeAxis.coordsystem
AXIS2STE + TimeAxis.accuracy.satatError.ErrorRefVal.ErrorRefValue
AXIS2SYE + TimeAxis.accuracy.sysError.ErrorRefVal.ErrorRefValue
AXIS2IND + TimeAxis.independentaxis
...
```

Ongoing work (2)

- **003.7858-39.2202.fits + MappingSpecificAxis.map ----> <u>003.7858-39.2202.uty</u>**

```
%CharacterisationAxis 1
%SpatialAxis.AxisName spatial
%SpatialAxis.independentaxis TRUE
%SpatialAxis.calibrationStatus CALIBRATED
%SpatialAxis.samplingPrecision.samplingPrecisionRefVal.samlingPeriod -0.000277777784317036
-0.000277777784317036
%SpatialAxis.coverage.bounds.limits.Coord2VecInterval.LoLimit2Vec 3.872320772806-39.08143766442968
%SpatialAxis.unit deg
%SpatialAxis.undersamplingStatus FALSE
%SpatialAxis.coordsystem FK5
%SpatialAxis.accuracy.statError.ErrorRefval.ErrorRefValue Unknown
%SpatialAxis.resolution.resolutionRefVal Unknown
%SpatialAxis.ucd pos
%SpatialAxis.numBins 512 1024
%SpatialAxis.regularsamplingStatus TRUE
%SpatialAxis.coverage.bounds.limits.Coord2VecInterval.HiLimit2Vec 3.762143519194-39.36588211557032
%SpatialAxis.accuracy.sysError.ErrorRefval.ErrorRefValue Unknown
%SpatialAxis.coverage.location.coord.Position2D.Value2.C1 3.8172321
%SpatialAxis.coverage.location.coord.Position2D.Value2.C2 -39.223659890

%CharacterisationAxis 2
%TimeAxis.AxisName time
%TimeAxis.coordsystem TT-ICRS-WAVELENGTH-TOPO
%TimeAxis.undersamplingStatus TRUE
%TimeAxis.numBins 1
%TimeAxis.accuracy.satatError.ErrorRefVal.ErrorRefValue Unknown
%TimeAxis.resolution.resolution.resolutionRefVal Unknown
...
```

- **Characterization library (VOTECH) is used to convert this format to an XML file**

Summary of this study

- **Done**
  - **Definition of workflow use cases with Characterized image entries**
  - **Definition of a constraint language and integration in AÏDA**
  - **Definition of constraint files for the use cases**
  - **...**
- **Ongoing work**
  - **Increase the validation scope**
    - During the execution : finalize the Characterization file generation for the FITS
    - Before the execution : study how to define a “virtual” Characterization file for an output before the execution...
    - Less human interaction
- **Full demo at next interop**