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.

Data Mining, J. Berthier et al.

TITAN/NOAR, L. Chevallier.

Simulation, F. Le Petit et al.

Trieste, IVOA meeting 19-23 May 2008
GWS Session 1, 20 May
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, ...)

![Workflow diagram](image)
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, ...

- ...

Trieste, IVOA meeting 19-23 May 2008
GWS Session 1, 20 May
Characterization in workflows
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
  - ...

Trieste, IVOA meeting 19-23 May 2008
GWS Session 1, 20 May
Characterization in workflows
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

<table>
<thead>
<tr>
<th>Tool</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Generic Char.</td>
</tr>
<tr>
<td>Constraints</td>
<td></td>
</tr>
<tr>
<td>Programme</td>
<td></td>
</tr>
</tbody>
</table>

#### 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

---

<table>
<thead>
<tr>
<th>Image Charac.</th>
<th>Entry₁</th>
<th>Description</th>
<th>Constraints</th>
<th>Programme</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image Charac.</td>
<td>Entry₂</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Trieste, IVOA meeting 19-23 May 2008
GWS Session 1, 20 May
Characterization in workflows
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 ...
Definition of the constraints

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)
ELSÈ
    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)
ELSÈ
    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

... and defined the grammar to generate the constraints parser

Very close to our needs (and to Characterization)
Third step: validation report generation
AÏDA client with validation capabilities
AÏDA client with validation capabilities (2)
Ongoing work

- **Characterization generation from FITS files, example:** 003.7858-39.2202.fits + MappingSpecificAxis.map ----> 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.samplingPeriod
  - AXIS2NAM + TimeAxis.AxisName
  - AXIS2UCD + TimeAxis.ucd
  - AXIS2UNI + TimeAxis.unit
  - AXIS2CAL + TimeAxis.calibrationStatus
  - AXIS2SYS + TimeAxis.coordsystem
  - AXIS2STE + TimeAxis.accuracy.statError.ErrorRefVal.ErrorRefValue
  - AXIS2SYE + TimeAxis.accuracy.sysError.ErrorRefVal.ErrorRefValue
  - AXIS2IND + TimeAxis.independentaxis
  ...

---

**Trieste, IVOA meeting 19-23 May 2008**

**GWS Session 1, 20 May**

**Characterization in workflows**
## Ongoing work (2)


```plaintext
%CharacterisationAxis 1
%SpatialAxis.AxisName spatial
%SpatialAxis.independentaxis TRUE
%SpatialAxis.calibrationStatus CALIBRATED
%SpatialAxis.samplingPrecision.samplingPrecisionRefVal.samplingPeriod -0.00277777784317036
-0.00277777784317036
%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.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.statError.ErrorRefVal.ErrorRefValue Unknown
%TimeAxis.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