Ensuring Long-term Preservation and Adding Value to Scientific and Technical Data

Olivier Rouchon – CINES

olivier.rouchon@cines.fr

PASIG 2012 – Friday, January 13th
**Centre Informatique National de l’Enseignement Supérieur**

Based in Montpellier, France

2 main mandates / activities:

- High performance computing
- Digital preservation
  - Electronic PhD theses
  - Digitized publications
  - Multimedia pedagogics
  - Scientific datasets

- Created in 1999, formerly known as **CNUSC** (Centre National Universitaire Sud de Calcul) – created in 1980
- Administered and funded by Ministry of Higher Education & Research (MESR)

January 13th, 2012

Austin PASIG 2012
Step 1: Source code, input data upload

Step 2: Data transfer for mid-term storage → added value

Jade is a SGI Altix ICE 8200 scalar, parallel supercomputer capable of a Rmax of 267 Tflops/s

Project = 1 year (renewable)
In 2011, 790 TB were generated from the 367 projects which have been hosted/run on the CINES HPC infrastructure.

This volume can be split over 10 thematic committees, among which:
- Environment
- Astrophysics and geophysics
- Fluid dynamics, reactive fluids, complex fluids
- Theoretical physics and plasmas
- Ordered molecular systems and biology

«Environment» represents 58% of the overall volume (450 TB)
Preservation challenges

Main risks deal with:
- Comprehension
- Integrity
- Exploitation
- Valorization

Quality assurance procedures to be setup for
- Metadata
- File formats
- Representation information
- Storage
- Access
- Technology watching
Setup quality assurance procedures to mitigate the impact of the four main identified risks when they occur

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loss of content knowledge</td>
<td>• Metadata; • Persistent, unique identifiers.</td>
</tr>
<tr>
<td>File format obsolescence</td>
<td>• Handling of a limited set of durable formats; • File format identification, validation; • Logical migration (format conversion).</td>
</tr>
<tr>
<td>Storage media failure</td>
<td>• Management of media ageing; • Physical migration.</td>
</tr>
<tr>
<td>Software or hardware disappearance</td>
<td>• Technology watching, anticipation, proactivity.</td>
</tr>
</tbody>
</table>
Preservation challenges

Can this be achieved with such volumes?

- Bytestream preservation not a problem
- Metadata and file formats are an issue

→ Need to evaluate the data which has a value and has to be preserved

→ Delete to preserve

Are researchers showing interest for HPC data preservation?

- Interest rate = 30% of overall projects
- Rate is identical within thematic committees.
- No correlation with project size
A project to preserve HPC data

• CINES will offer in 2012 "mid-term preservation services" for scientific data for the attention of interested user communities / thematic committees.

  – 3 to 4 years preservation/archive;
  – ISAAC Project – Information Scientifique Archivée Au CINES.

➔ The objectives are to

  – Give additional time for the researcher to appraise the relevance / importance of the information;
  – Put in place processes for scientific data valorization / preservation.

➔ This goes well beyond a simple storage or backup!

• At the end of this 3-4 years period, two options

  – Migration onto the long-term preservation platform – PAC;
  – Restitution to the producer/owner.
## Key features

<table>
<thead>
<tr>
<th>Functionalities</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Context / content knowledge</td>
<td>• Descriptive metadata (qualified DublinCore, METS);</td>
</tr>
<tr>
<td>• Four levels: community, producer, object, file;</td>
<td></td>
</tr>
<tr>
<td>• Use of community standards for metadata.</td>
<td></td>
</tr>
<tr>
<td>Information integrity</td>
<td>• Unique persistent identifiers for objects;</td>
</tr>
<tr>
<td>• Identification, validation of accepted formats (HDF5, etc.);</td>
<td></td>
</tr>
<tr>
<td>• Fixity checks;</td>
<td>• Multiple copies;</td>
</tr>
<tr>
<td>• Events logging and archiving.</td>
<td></td>
</tr>
<tr>
<td>Information valorization</td>
<td>• Representation information (thesis, papers, etc.)</td>
</tr>
<tr>
<td>Sharing / access</td>
<td>• Web-based user interface;</td>
</tr>
<tr>
<td>• Search on metadata;</td>
<td>• User authentication;</td>
</tr>
<tr>
<td>• Access rights management.</td>
<td></td>
</tr>
</tbody>
</table>

→ Management of producers and communities requests for improvement.
Representation information added value

Users or Community Knowledge base

Digital object

Interpreted using

UTF-8, HDF-5, NetCDF

Interpreted using

Comprehensible information

Article, Dictionary, Thesis

010100 110110 011

January 13th, 2012

Austin PASIG 2012
<table>
<thead>
<tr>
<th>Long-term preservation</th>
<th>High performance computing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Competencies:</td>
<td>Competencies:</td>
</tr>
<tr>
<td>• Archivists;</td>
<td>• Support of parallel</td>
</tr>
<tr>
<td>• File format experts.</td>
<td>computing codes;</td>
</tr>
<tr>
<td></td>
<td>• Storage of large volumes;</td>
</tr>
<tr>
<td></td>
<td>• Scientific data file</td>
</tr>
<tr>
<td>6 years experience in</td>
<td>formats.</td>
</tr>
<tr>
<td>service delivery</td>
<td></td>
</tr>
<tr>
<td>Tools:</td>
<td></td>
</tr>
<tr>
<td>• Applications: file</td>
<td></td>
</tr>
<tr>
<td>format identification</td>
<td></td>
</tr>
<tr>
<td>and characterization;</td>
<td></td>
</tr>
<tr>
<td>• Concepts: OAIS model</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Infrastructure**

Servers, tape libraries, replication, RENATER & GEANT networks.

[Logos: BnF, Renater, GÉANT]
A generic platform with a national/european scope:

- A prototype is being put in place for the PRECCINSTA (Prediction and control of combustion instabilities for industrial gas turbines) datasets – potentially 2TB of data – due June 2012;
- Developments based on standard, open technologies: Java, PostgreSQL, iRODS, OpenLDAP, etc. which can be reused;
- Partnerships with other european institutions as part of EC funded project: EUDAT (ReDS – Researcher Data Store);
- ISAAC willing to become the French node of a european data grid.

A shared environment of ideas, means and infrastructure:

- Pooling of projects on the platform, generic ingest/upload process;
- Reduction of implementation and operation costs.

A double competency on data and preservation:

- Dedicated department for preservation;
- Privileged relationship with researchers.
DATA USER

- Producer
- Authorized user
- Communities

Storage abstract Layer

Ingest API
- Format validation
- Metadata management
- Integrity checks
- Rights management

WEB INTERFACE

STORAGE

OpenLDAP

JAVA

XML

PostgreSQL
New challenges to take up:

- Raise awareness and involve users of scientific communities;
- Encourage/promote new metadata, file formats standard for scientific communities;
- Offer services in sync with users needs to generate support;
- Build an attractive business/economic model;
- Take at higher level – European grid for data.
Thanks for your attention

More information: http://www.cines.fr/
olivier.rouchon@cines.fr
Using loads of computational power in your research? Backing up all of your data just in case? You’re part of the problem! Become part of the solution…

Delete a petabyte

Save a polar bear

Join the campaign – delete some data today. You know you want to.

This message brought to you by Concerned Researchers for a Greener Planet.