The importance of a global approach for scientific data in the ERA building

Dany VANDROMME

Dany.Vandromme@renater.fr
www.renater.fr

Global approach to scientific data

- The size problem
- The organisational scheme
- The access, interoperability and usability....
- Role of users...

The size problem

- The most conventional example:
  - LHC: 10-15 Pbytes per year
- To-morrow challenges:
  - Biology : in the range of Pbytes, but widely heterogeneous and distributed
  - Astronomy: SKA claims to produce 100 times LHC figures
  - All other disciplines (not listed here) racing for highest prospectives...

The size problem

- A fact: Volume of data is growing (almost out of control)
- Data are produced everywhere and used everywhere
- Two trends to face:
  - Accompany the growth: efficiency, availability
  - Limit the growth: curation, green issue

The organizational issue

- Back to the classic LHC:
  - Data production at CERN
  - First level of treatment (filtering) at CERN → T0
  - Distributed storage over the T1 (EU, US and TW)
  - Access, usability and computing: Fully distributed and shared resources (T2, T3)

The organizational issue

- Two ways to justify the data architecture:
  - Too many data to be managed centrally (cost, network, technicality issues)
  - CERN is a member organization: Members should bear (contribute) the load, outside the CERN budget...
  - The users (physicists) are all around the world: Data should be deliver to them wherever they are.

The organizational issue

- Another example: VLBI (EU or global)
  - Data (images) are produced by radiotelescopes and shipped to a correlator (in NL) which produce the synthesis.
  - Not scalable at the global level as the correlation should (and will) be also a distributed process. Then all worldwide antenna will “correlate”.
  - Network bandwidth is not yet fully solved for the future….

The organizational issue

- A third example: HPC (DEISA, PRACE, etc...)
  - These machines will produce more data than any LHC
  - The present vision is to have large storage near the CPUs and eventually users accessing remotely
  - LHC-type global architecture model is still to be worked out

The organizational issue

- A third example: HPC (DEISA, PRACE, etc...) (continued)
  - So far, grid systems (as computing resources) did not meet this difficulty as they are totally distributed and happy with Internet-type networks
  - Future of HPC will require more taylored architecture (still a missing part of DEISA, but what about PRACE?)

The access, interoperability and usability….

- The critical issue: Various initiatives aims to contribute to the problem and possible solutions: EC communications, ESFRI statements and WG, OECD contribution, E-IRG DMTF, EC funding etc...
- Aim is to build a scientific data infrastructure!

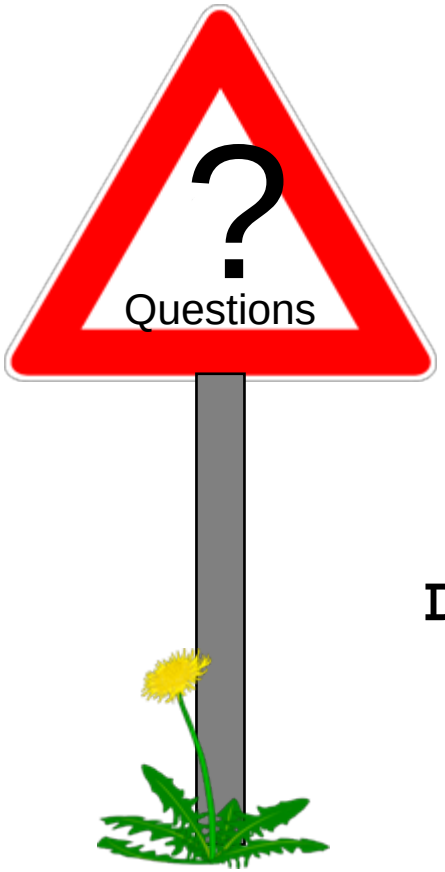
The access, interoperability and usability….

- To transport is not a real issue, but to make it efficiently for ad-hoc use may turn to be quite complicated for a number of scientific disciplines.
- Physics looks very comfortable compared to SSH or LS
- Environment is a different context as it requires also complete merging with EO data, modelling etc...

Now the down side of the growth

- Curation: We cannot afford to store and maintain all data! But we have not yet the proper criteria to curate...
- Green aspects: Not yet a real issue today, except for HPC or massive concentrated storage centers (like Google ☺), but think about to-morrow (5-10 years from now….)

End of presentation



Dany.Vandromme@renater.fr