



Enabling Grids for E-scienceE

Biomed Data Challenge

data collect

Cyril L'Orphelin, Geneviève Romier
Pisa, Tuesday, 25 October 2005

www.eu-egee.org



- **Collect all data concerning the Biomed Data Challenge jobs on EGEE during the data challenge period (summer 2005)**
- **Provide Biomed team with usual and specific statistics**
- **This Data challenge is also a way to validate our results on a representative and well identified sample of jobs.**

- **Several weeks before the beginning**
 - get information from data challenge team
 - we agreed about the RBs foreseen to be used (among the RBs collected)
 - JRA2 statistics people organized his summer vacation to follow up the data challenge as much as possible
- **During the data challenge**
 - adapt the method to the DC reality
 - DC team needed to use other RBs
 - *ask access rights to collect data from with the efficient help of the DC team*
 - follow up the data collect all the DC duration
- **After the data challenge**
 - get information from data challenge team
 - to select and identify all jobs concerned among all data before to provide statistics
 - to provide the Biomed people with the statistics they asked for.

- **according to RBs**
 - new RBs to be collected
 - we had to build a filter to collect only Biomed data for this period for a RB whose administrator did not accept to give us access
- **according to data**
 - select and identify the DC data among Biomed data
 - we need clear, accurate and usable criteria (UI, RBs, users...)
 - the work was almost manual
- **according to statistics**
 - new scripts for specific statistics needed by Wisdom community

- **The team work is essential prior to the DC, during the DC and after the DC**
- **The selection criteria must be discussed and defined before the DC.**
- **It was an occasion to experiment new statistics (e.g. global duration)**
- **We need for the future the application tag in the job history record to be used as a criterium to identify the DC jobs**