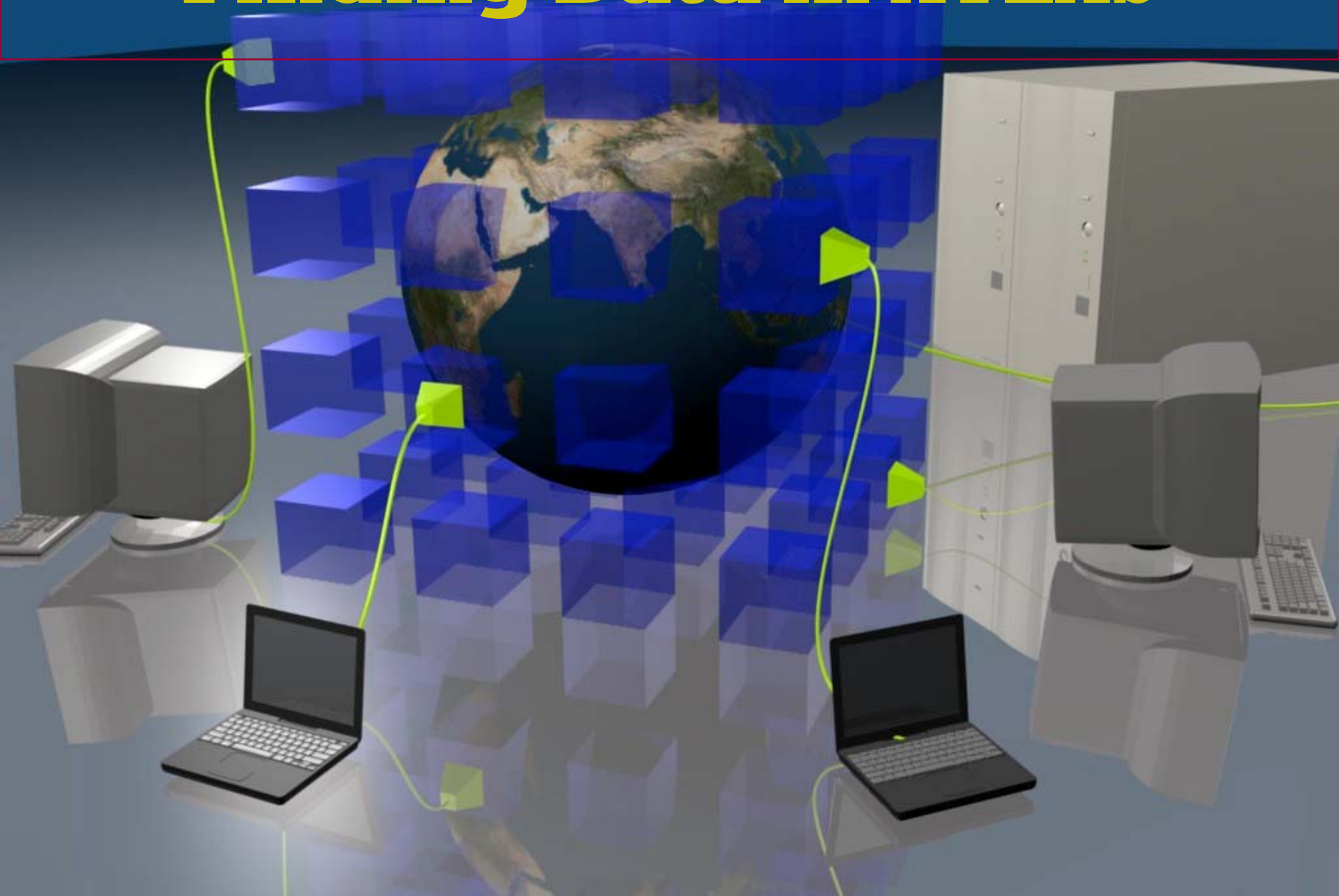
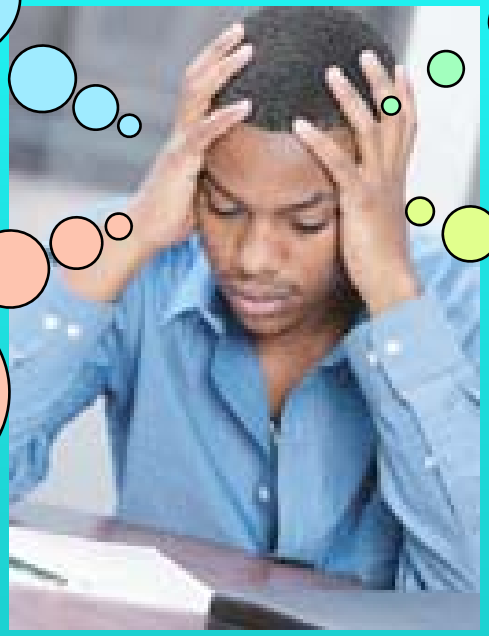
Finding Data in ATLAS



Starting Point Questions

Are there are any AOD produced with release 15?

What triggers were used for this dataset?



What geometry do I need to analyze this data?

What is the latest reprocessing of cosmics?

Focus of this Talk

- Although all of these questions are relevant, this talk will concentrate on exploring the data store and ***not*** on configuring a job to use the data store as input.
  - The exploration will nevertheless demonstrate how to find information that ***will*** be used to configure the job.

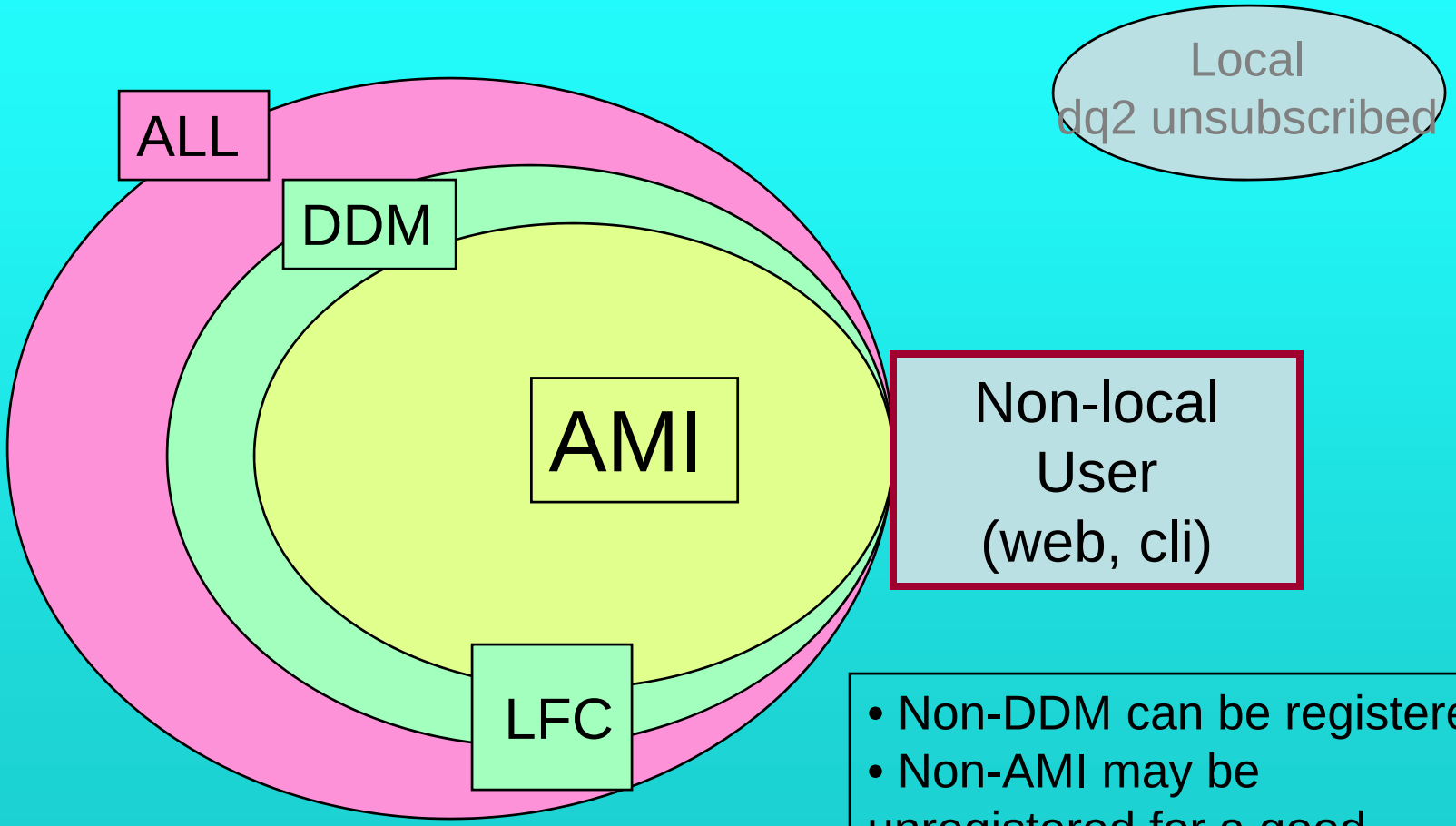
Jargon and Definitions

- A Dataset is
  - a group of events (Event Store, Streams)
  - a group of files (DDM)
  - something registered as a 'dataset' (AMI)
- GUID = alphanumeric unique file identifier
- LFN = logical file name
  - An alias unconnected to actual storage,
    - e.g. MyEvents.AOD.rel14
  - Whatever is used in the LFN field in the LFC
    - E.g. /data/59222.cosmics_nomag/AOD/00001.file.1
- PFN = physical file name
  - A pointer to a physical instance of a file
    - dcache:/pnfs/panda/data/data09. 59222.cosmics.AOD.00001.file.1
- LFC = LCG File Catalog which contains
  - GUID's, LFN's, PFN's, (file metadata)
  - Local to a site

(Cont.)

- AMI = ATLAS Metadata Interface
  - Tool for browsing metadata collected in DB's at CERN/Lyon.
- DDM = Distributed Data Management
  - Groups files into file datasets
  - Has one level of hierarchy called "containers"
  - Information for worldwide distribution is tracked in central catalogs at CERN
- DQ2 (Don Quijote)
  - General set of tools used to transfer files tracked by the DDM catalogs
    - dq2-ls, dq2-get, …
- PanDA = Production and Distributed Analysis
  - System for running jobs on ATLAS computing resources and the grid
  - Uses DDM to manage input and output files.

Data Visibility



- Non-DDM can be registered.
- Non-AMI may be unregistered for a good reason.

Stage I: Identifying Datasets

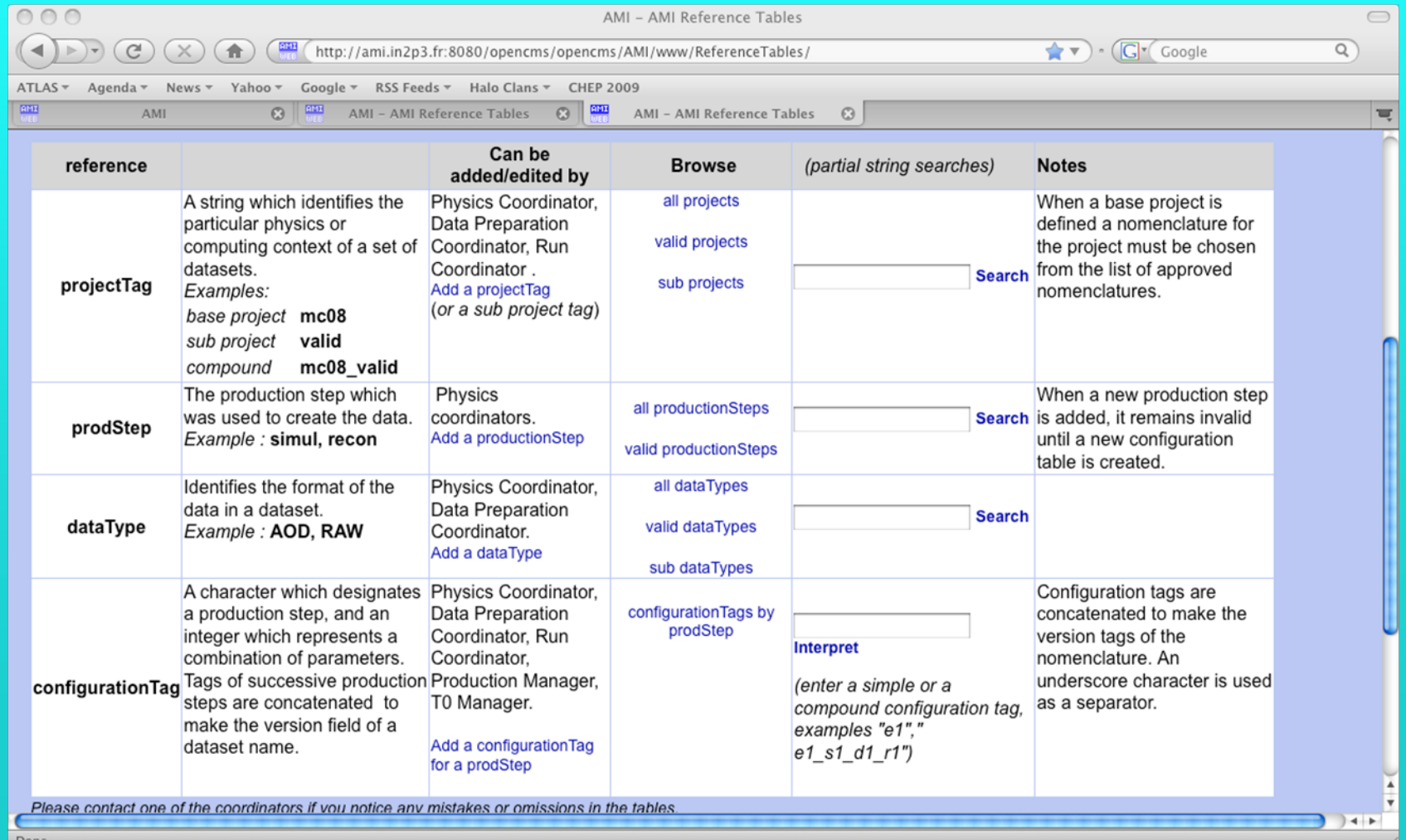
- How is the ATLAS data organized at a macro level?
  - Organized into projects
  - Projects can be specialized in what metadata they use or track.
    - E.g. Monte Carlo would track generators while real data would not. These are in different projects.
  - Other metadata at the macro level
    - prodStep, e.g. recon
    - dataType, e.g. AOD
    - configurationTag ('AMI tag'), e.g. o4_r653_p27
- A good place to start is AMI
  - http://ami.in2p3.fr/opencms/opencms/AMI/www/

Searches in AMI

- http://ami.in2p3.fr/opencms/opencms/AMI/www/
- Back button can be tricky, be prepared to pulse it.
- Reference Tables
  - Macro data, orientation
- Overview
  - See what data is available
  - Get a handle on variety and volumes
- Advanced
  - More easily choose things like release or data type (AOD, DPD, etc.)
- Simple
  - You have a clear idea of how the system works and you just need to do a wildcard search.
    - E.g. mc08%Higgs%ESD.15%

Macro Data

- Getting oriented, let's browse the macro data.
- http://ami.in2p3.fr:8080/opencms/opencms/AMI/www/ReferenceTables/



reference		Can be added/edited by	Browse	(partial string searches)	Notes
projectTag	A string which identifies the particular physics or computing context of a set of datasets. *Examples:* *base project* **mc08** *sub project* **valid** *compound* **mc08_valid**	Physics Coordinator, Data Preparation Coordinator, Run Coordinator . Add a projectTag (*or a sub project tag*)	all projects valid projects sub projects	Search	When a base project is defined a nomenclature for the project must be chosen from the list of approved nomenclatures.
prodStep	The production step which was used to create the data. *Example :* **simul, recon**	Physics coordinators. Add a productionStep	all productionSteps valid productionSteps	Search	When a new production step is added, it remains invalid until a new configuration table is created.
dataType	Identifies the format of the data in a dataset. *Example :* **AOD, RAW**	Physics Coordinator, Data Preparation Coordinator. Add a dataType	all dataTypes valid dataTypes sub dataTypes	Search	
configurationTag	A character which designates a production step, and an integer which represents a combination of parameters. Tags of successive production steps are concatenated to make the version field of a dataset name.	Physics Coordinator, Data Preparation Coordinator, Run Coordinator, Production Manager, T0 Manager. Add a configurationTag for a prodStep	configurationTags by prodStep	Interpret *(enter a simple or a compound configuration tag, examples "e1", "e1_s1_d1_r1")*	Configuration tags are concatenated to make the version tags of the nomenclature. An underscore character is used as a separator.

Please contact one of the coordinators if you notice any mistakes or omissions in the tables.

AMI Portal Home
Dataset Search Tutorial
Client
Developer
Presentations
Principles of AMI Design
Nomenclature
Metadata Dictionary

Welcome to the ATLAS Metadata Interface!

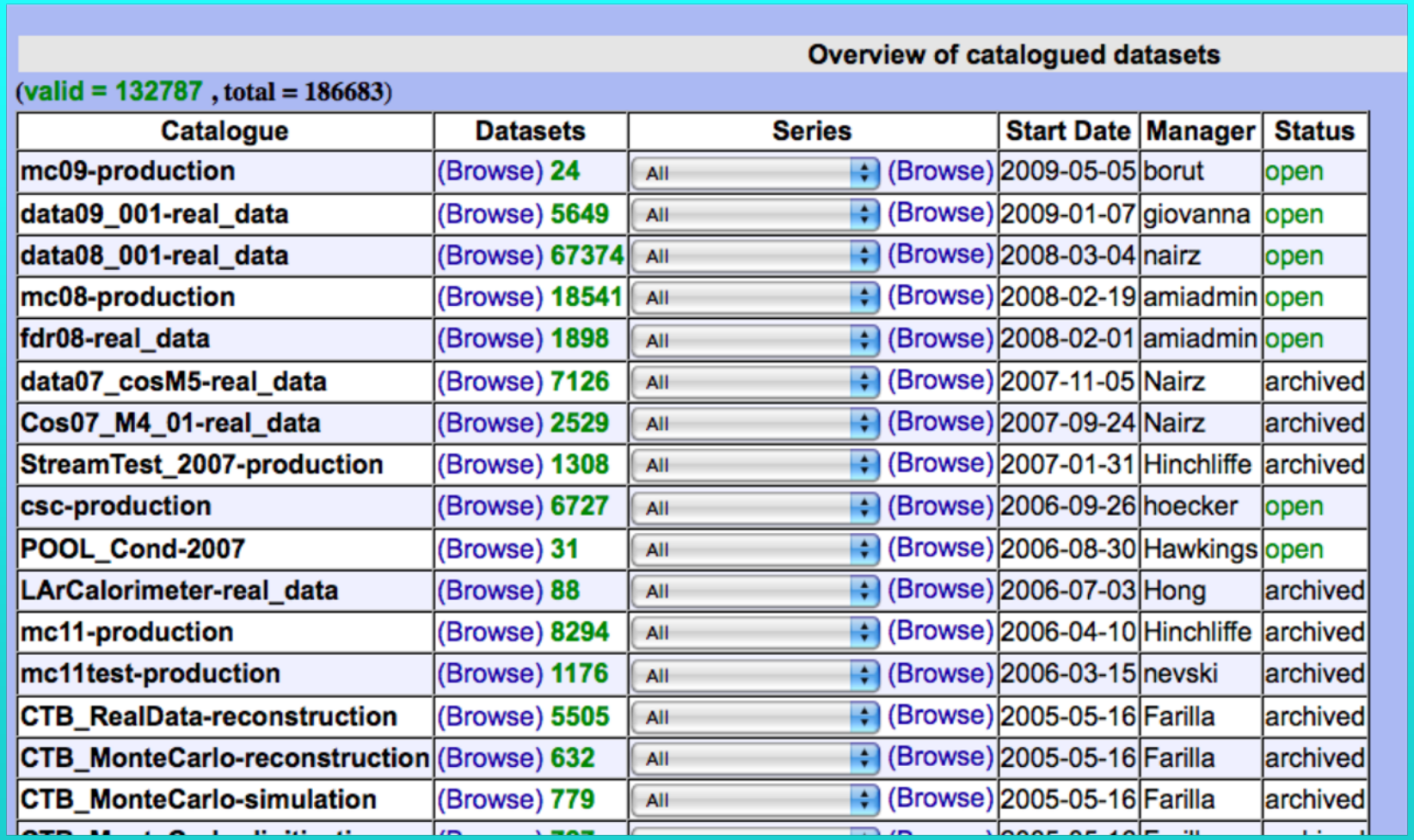
*This page gives access to two ATLAS tools (choose the **https** links to authenticate with your certificate):-*

The Tag Collector for ATLAS software release management. http, https

The AMI Dataset Search of ATLAS real and simulated data. http, https

- An overview of all the datasets catalogued in AMI. (http, https)
- The simplest way of searching is by name (fast) or keyword (longer, because many fields are searched).(http,https) links to the dataset search and the configuration tag interpretation
- The advanced search lets you set search criteria on some selected fields. N.B. By default AMI hides datasets which are deleted or known to be bad, this can be disabled when you use the advanced search. (http, https)
- Interpretation of the dataset configuration tags. (http, https)
- Once you get your results, you can refine them, either by using the selection functions attached to the columns of your result set, or by going to the powerful "Refine Query " interface. If you are new to databases and SQL, we advise you to work through the tutorial before using the "Refine Query" functions. We can also provide bookmarks to complex queries on demand.

Getting an Overview



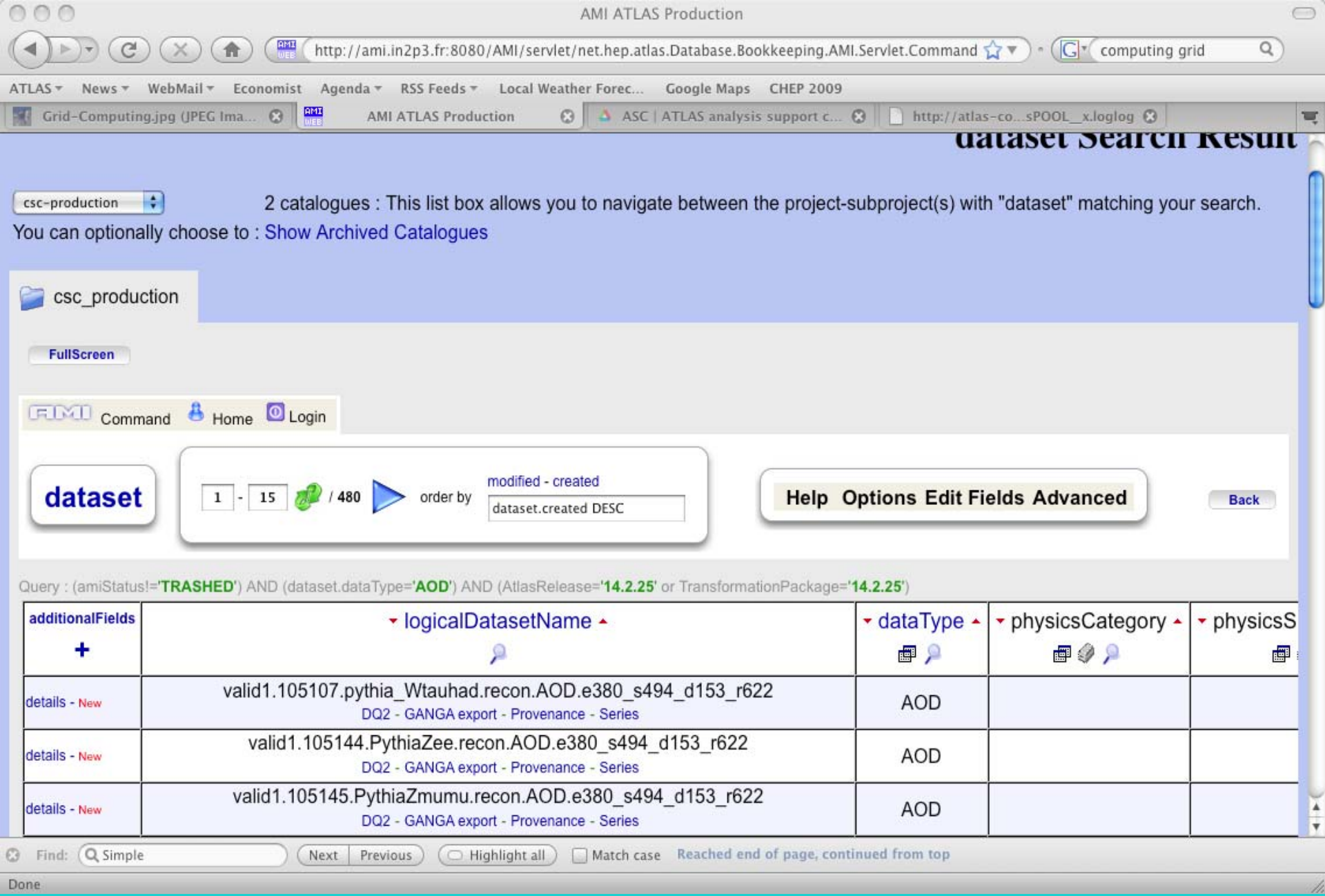
Overview of catalogued datasets

(valid = 132787 , total = 186683)

Catalogue	Datasets	Series	Start Date	Manager	Status
mc09-production	(Browse) 24	All (Browse)	2009-05-05	borut	open
data09_001-real_data	(Browse) 5649	All (Browse)	2009-01-07	giovanna	open
data08_001-real_data	(Browse) 67374	All (Browse)	2008-03-04	nairz	open
mc08-production	(Browse) 18541	All (Browse)	2008-02-19	amiadmin	open
fdr08-real_data	(Browse) 1898	All (Browse)	2008-02-01	amiadmin	open
data07_cosM5-real_data	(Browse) 7126	All (Browse)	2007-11-05	Nairz	archived
Cos07_M4_01-real_data	(Browse) 2529	All (Browse)	2007-09-24	Nairz	archived
StreamTest_2007-production	(Browse) 1308	All (Browse)	2007-01-31	Hinchliffe	archived
csc-production	(Browse) 6727	All (Browse)	2006-09-26	hoecker	open
POOL_Cond-2007	(Browse) 31	All (Browse)	2006-08-30	Hawkings	open
LArCalorimeter-real_data	(Browse) 88	All (Browse)	2006-07-03	Hong	archived
mc11-production	(Browse) 8294	All (Browse)	2006-04-10	Hinchliffe	archived
mc11test-production	(Browse) 1176	All (Browse)	2006-03-15	nevski	archived
CTB_RealData-reconstruction	(Browse) 5505	All (Browse)	2005-05-16	Farilla	archived
CTB_MonteCarlo-reconstruction	(Browse) 632	All (Browse)	2005-05-16	Farilla	archived
CTB_MonteCarlo-simulation	(Browse) 779	All (Browse)	2005-05-16	Farilla	archived

Drill Down Options

- Find transform and parameters used by the jobs which created a dataset.
- Find which release, geometry, calibration were used.
- Find sites which are subscribed to this dataset.
- Find information on the run configuration for DAQ runs.
- Find the datasets from which a dataset was derived and their parameters.
- …


AMI ATLAS Production
http://ami.in2p3.fr:8080/AMI/servlet/net.hep.atlas.Database.Bookkeeping.AMI.Servlet.Command
computing grid
ATLAS News WebMail Economist Agenda RSS Feeds Local Weather Forec... Google Maps CHEP 2009
Grid-Computing.jpg (JPEG Ima...
AMI ATLAS Production
ASC | ATLAS analysis support c...
http://atlas-co...sPOOL_x.loglog
dataset Search Result
csc-production
2 catalogues : This list box allows you to navigate between the project-subproject(s) with "dataset" matching your search.
You can optionally choose to : Show Archived Catalogues
csc_production
FullScreen
AMI Command Home Login
dataset
1 - 15 / 480
order by
modified - created
dataset.created DESC
Help Options Edit Fields Advanced
Back
Query : (amiStatus!='TRASHED') AND (dataset.dataType='AOD') AND (AtlasRelease='14.2.25' or TransformationPackage='14.2.25')
additionalFields
logicalDatasetName
dataType
physicsCategory
physicsS
details - New
valid1.105107.pythia_Wtauhad.recon.AOD.e380_s494_d153_r622
DQ2 - GANGA export - Provenance - Series
AOD
details - New
valid1.105144.PythiaZee.recon.AOD.e380_s494_d153_r622
DQ2 - GANGA export - Provenance - Series
AOD
details - New
valid1.105145.PythiaZmumu.recon.AOD.e380_s494_d153_r622
DQ2 - GANGA export - Provenance - Series
AOD
Find: Simple Next Previous Highlight all Match case Reached end of page, continued from top
Done

Questions You Might Ask

- How do I know if a filter was applied at a certain stage in the provenance chain?
  - Corollary: What was the event loss at each processing stage and why?
  - AMI team working to collect this.
- How do I know what Container#Key combos are available in the datasets?
  - Corollary: Do all events have the same content?
  - Look at a file and assume all are the same.
- Are there any differences in content between AOD in 14.2.22.5 and 14.5.1?
  - Corollary: Can I run the same job on datasets from both of those releases?
  - Assume if the first two decimals are the same that it is yes, otherwise no, or try it and see.

Stage 2: Finding a Copy

- AMI has allowed you to identify datasets. Now you'd like to do something with that data.
  - Run an analysis on the full dataset.
  - Pull over a small portion to test.
  - Extract a subset based on some criteria.
  - Make some plots of data content
  - …
- Now is where we enter the DDM realm.

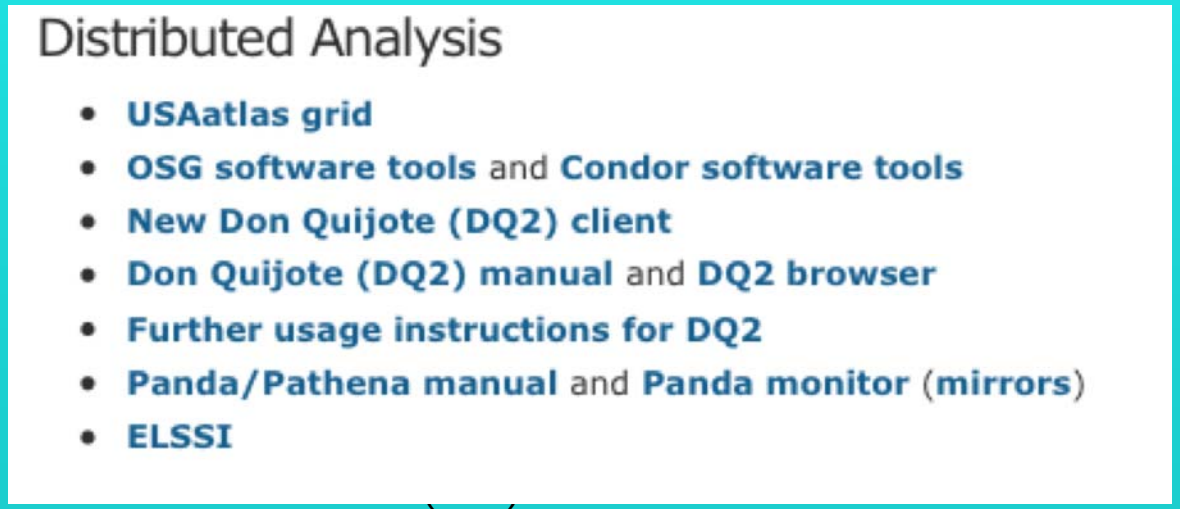
DDM: The Layer Above

- All official ATLAS data is managed by DDM and distributed based on the "Tiers of ATLAS".
  - Tier 0: RAW, first reco, calib
  - Tier 1: partRAW,partESD,AOD,reproc,DPD
  - Tier 2: partESD, partAOD, DPD, simul
- Distributions are managed by subscriptions
- Subscribing takes some privileges
- More info at DDMOperationsGroup twiki.

DQ2: command line tools

- https://twiki.cern.ch/twiki/bin/view/Atlas/DQ2ClientsHowTo
  - Reasonably comprehensive set of definitions and examples. Should get you started.
  - Includes setup for CERN
  - Setup by default at ASC

ASC "Useful Links"



Eggsample (1)

- Finding "container" datasets (datasets of datasets) (~3 min)

```
lxplus% time dq2-ls data08*.TAG*/
data08_cosmag.00092072.physics_L1Calo.recon.TAG_COMM.o4_r653/
data08_cos.00092048.physics_L1CaloEM.merge.TAG_COMM.o4_r653_p27/
…
data08_cosmag.00092092.physics_L1Calo.merge.TAG.o4_f74_m34/
data08_cosmag.00091560.physics_IDCosmic.merge.TAG_COMM.o4_f71_m19/
data08_cosmag.00090801.physics_TGCwBeam.recon.TAG_COMM.o4_r653/
```

  - 5407 lines

- Contents of a container dataset (tid's)

```
dq2-list-datasets-container
   data08_cosmag.00091387.physics_IDCosmic.merge.TAG_COMM.o4_r653_p27/
    data08_cosmag.00091387.physics_IDCosmic.merge.TAG_COMM.o4_r653_p27_tid062482
    data08_cosmag.00091387.physics_IDCosmic.merge.TAG_COMM.o4_r653_p27_tid062483
    …
```

Eggsample(2)

- Finding a container replica (~5 min)

```
lxplus% time dq2-ls -r data08_cosmag.00091387.physics_IDCosmic.merge.TAG_COMM.o4_r653_p27/

data08_cosmag.00091387.physics_IDCosmic.merge.TAG_COMM.o4_r653_p27/
 No replicas available!

lxplus% dq2-ls -r data08_cosmag.*.physics_IDCosmic.merge.AOD*

…
data08_cosmag.00090262.physics_IDCosmic.merge.AOD.o4_r602_p16_tid033292
 INCOMPLETE:
 COMPLETE:
   BNLPANDA
   FZK-LCG2_DATADISK
   RAL-LCG2_DATADISK
   INFN-T1_DATADISK
   TRIUMF-LCG2_DATADISK
   NDGF-T1_DATADISK
   SARA-MATRIX_DATADISK
   PIC_DATADISK
   CERN-PROD_DATADISK
   IN2P3-CC_DATADISK
```

Eggsample(3)

- Get a dataset for local use

```
lxplus% dq2-get
  data08_cosmag.00091387.physics_IDCosmic.merge.TAG_COMM.o4_r6
  53_p27/
…
lxplus% ls
data08_cosmag.00091387.physics_IDCosmic.merge.TAG_COMM.o4_r653_p27_tid062523
data08_cosmag.00091387.physics_IDCosmic.merge.TAG_COMM.o4_r653_p27_tid062524
data08_cosmag.00091387.physics_IDCosmic.merge.TAG_COMM.o4_r653_p27_tid062525
data08_cosmag.00091387.physics_IDCosmic.merge.TAG_COMM.o4_r653_p27_tid062526
data08_cosmag.00091387.physics_IDCosmic.merge.TAG_COMM.o4_r653_p27_tid062527
lxplus% ls | wc
     76      76    5852
```

DQ2: web interface

- Currently under construction.
- There are still discontinuities in the system. Do cross checks!
  - Does number of files delivered match files shown by dq2-ls?
  - If there are no replicas, does a dq2-get still work?
  - Talk to people, e.g. Contact the person in charge of that project or production with results.
  - …

Conclusion

- The good
  - There are web and cli tools which allow you to browse and find data.
  - The documentation does have lots of examples for some common use cases.
- The bad
  - Some searches can be slow (>10min).
  - Some information is incomplete.
- The ugly
  - When the info is available, the presentation or interface (click) may be uncomfortable.