Managing Data

DIRAC Tutorial

Outline

- Data management components
  - Storage Elements
  - File Catalogs
- DIRAC conventions for user data
- Data operation commands
- Data bookkeeping with the File Catalog CLI
  - Replica Catalog
  - Metadata Catalog

Data Management components

Storage Elements

- gLite/EGI Storage Elements
  - Standard SRM interface
  - Gridftp protocol
    - Need Globus libraries, limited number of platforms
  - Allow third party transfers between them
  - Managed by the site managers within EGI SLAs
- DIRAC Storage Elements
  - DISET based components
  - DIPS (Dirac Secure Protocol)
  - Does not allow third party transfers
    - Replication through local cache
    - Third party transfers will be available in the future
- Irods Storage Elements

Data Management components

- File Catalogs
  - LCG File Catalog (LFC)
    - Part of the EGI middleware
    - Service provided by the NGI
      - ORACLE backend
    - Client tools: command line, Python API
      - Need Globus libraries
    - No User Metadata support
  - DIRAC File Catalog
    - DISET based components
    - Part of the DIRAC set of services
      - Community service
      - MySQL backend
    - Client tools: command line, CLI, Python API
    - Support of the User Metadata

Data Management components

- For DIRAC users the use of any Storage Element or File Catalog is transparent
  - Community choice which components to use
  - Different SE types can be mixed together
  - Several File Catalogs can be used in parallel
    - Complementary functionality
    - Redundancy
- Users see depending on the DIRAC Configuration
  - Logical Storage Elements
    - e.g. DIRAC-USER, M3PEC-disk, MCIA-irods
  - Logical File Catalog

DIRAC data naming conventions

Each file is identified by its Logical File Name (LFN)

- Primary unique identifier
- GUIDs are supported but their uniqueness is under the responsibility of user applications
  - This is different from LFC
  - Mostly for support of some applications, e.g. ROOT I/O

LFN construction

- Starts always with the VO name
  - /vo.formation.idgrilles.fr/…
- User data
  - /vo.formation.idgrilles.fr/user/v/vhamar/…

PFN (Physical File Name) construction

- Always contains LFN as it trailing part

Data operation commands

- **`dirac-dms-add-file`**
  - Upload file to the grid SE (lcg-cr)
- **`dirac-dms-get-file`**
  - Download file to the grid SE (lcg-cp)
- **`dirac-dms-replicate-lfn`**
  - Make another replica of a file (lcg-rep)
- **`dirac-dms-lfn-replicas`**
  - List replicas of a given file (lcg-lr)
- **`dirac-dms-user-lfns`**
  - Get a list of all the user files
- **`dirac-dms-find-lfns`**
  - Find a list of all the user files corresponding to a given search criteria
- Plus others …
  - See tutorial materials

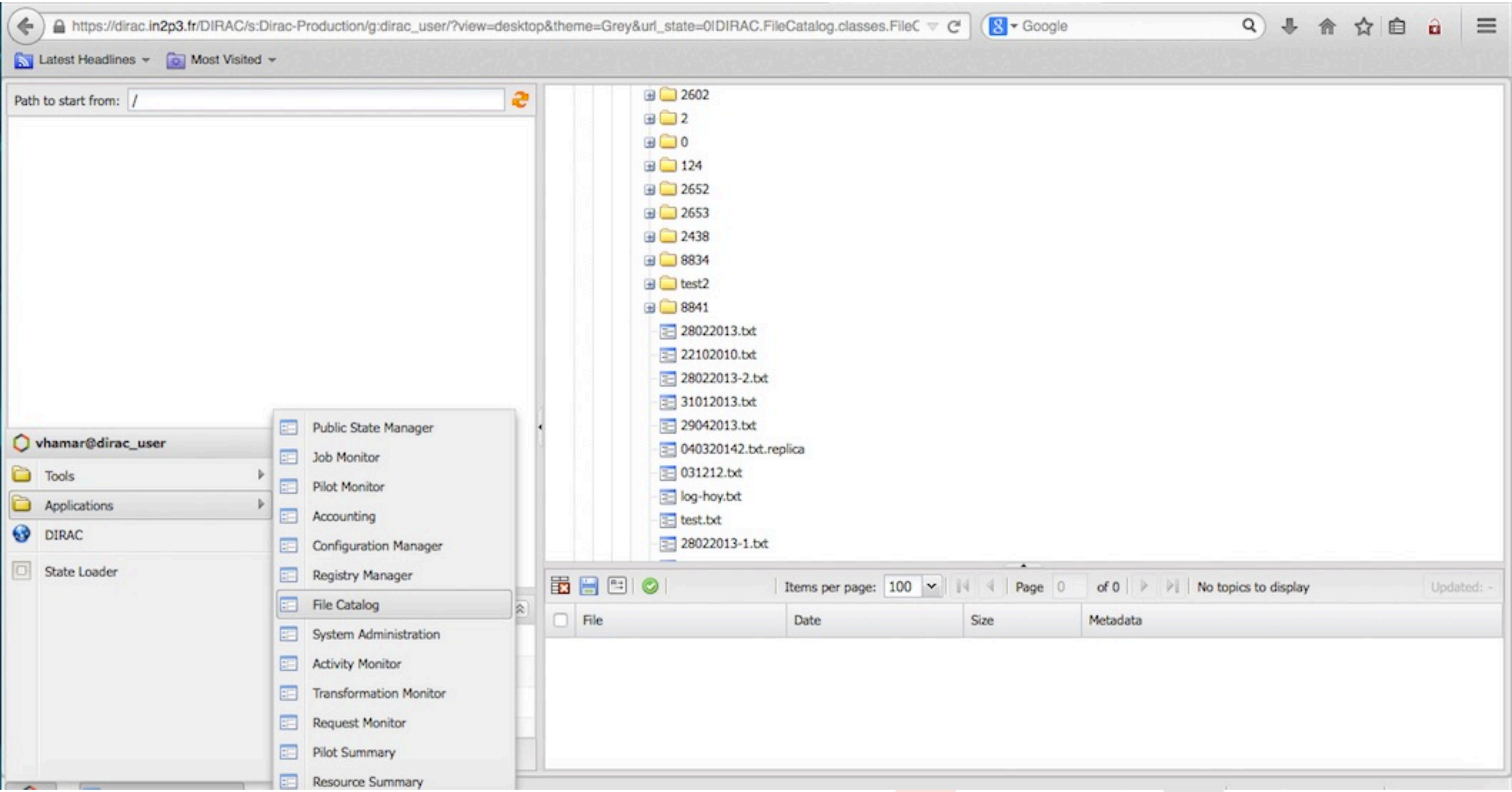
COMDIRAC
Data Management Commands

- **`dput`** - works like a kind of scp command (in the local->remote direction):
  - `dput <local-source-path> [<lfn-destination-file-or-directory>]`
  - `dput <local-source-path>... <lfn-destination-directory>`
- **`dget`** - works in the opposite direction (remote->local):
  - `dget <lfn-source-path>`
  - `dget <lfn-source-path>... <local-destination-directory>`

COMDIRAC
File Catalog Commands

- **`dpwd`** – shows current directory
- **`dcd`** – change directory
- **`dls`** - list FC directory contents
- **`dmkdir`** - create directory
- **`drmdir`** - remove directory
- **`dreplicas`** – file replication
- **`drm`** – file and directory removal
- **`dfind`** – find files with requested metadata

DIRAC Web Portal File Catalog



Tutorial

Usefull links

Follow this tutorial:

https://github.com/DIRACGrid/COMDIRAC/wiki/Data-Management

Other DIRAC data management tools

https://github.com/DIRACGrid/DIRAC/wiki/DataManagement

https://github.com/DIRACGrid/DIRAC/wiki/FileCatalog

With DIRAC command line tools

- Getting data files to the grid
- Downloading data files from the grid
- Replicating files
- Exploring the File Catalog console

Backup slides

Asynchronous operations

- File Catalog operations are generally synchronous
  - Quick, can wait for the prompt
- Physical data operations can take very long time
  - And even fail in the end
- For example, consider removing data:
  - Delete replicas on all the SEs
  - Delete files (lfns)
  - Delete directories (recursively)
- Long operations are performed asynchronously
  - Do not wait for completion
  - Make sure the operation is accomplished despite possible problems

File Catalog Metadata

Metadata can be associated with each directory as key:value pairs to describe its contents

- Int, Float, String, DateTime value types

Some metadata variables can be declared indices

- Those can be used for data selections

Subdirectories are inheriting the metadata of their parents

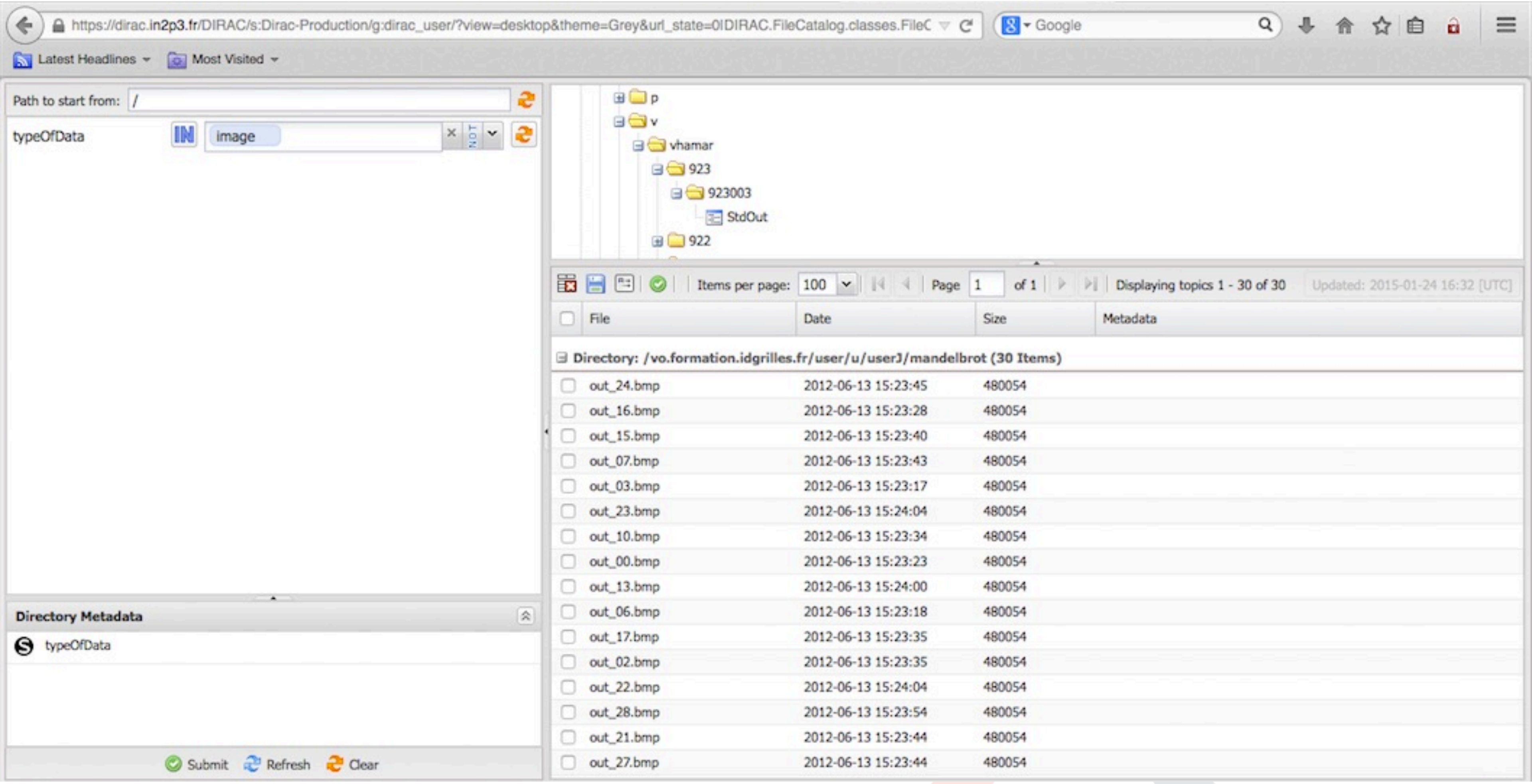
Data selection with metadata queries

- Example:

```
find /dirac/user Meta1=Value1 Meta2>3 Meta2<5 Meta3=2,3,4
```

File metadata can also be defined

File Catalog



https://dirac.in2p3.fr/DIRAC/s:Dirac-Production/g:dirac_user/?view=desktop&theme=Grey&url_state=0|DIRAC.FileCatalog.classes.FileC

Path to start from: /

typeOfData IN image

Directory Metadata: typeOfData

```
⊞ p
⊟ v
  ⊟ vhamar
    ⊟ 923
      ⊟ 923003
          StdOut
    ⊞ 922
```

Items per page: 100 | Page 1 of 1 | Displaying topics 1 - 30 of 30 | Updated: 2015-01-24 16:32 [UTC]

Directory: /vo.formation.idgrilles.fr/user/u/user3/mandelbrot (30 Items)

File	Date	Size	Metadata
out_24.bmp	2012-06-13 15:23:45	480054	
out_16.bmp	2012-06-13 15:23:28	480054	
out_15.bmp	2012-06-13 15:23:40	480054	
out_07.bmp	2012-06-13 15:23:43	480054	
out_03.bmp	2012-06-13 15:23:17	480054	
out_23.bmp	2012-06-13 15:24:04	480054	
out_10.bmp	2012-06-13 15:23:34	480054	
out_00.bmp	2012-06-13 15:23:23	480054	
out_13.bmp	2012-06-13 15:24:00	480054	
out_06.bmp	2012-06-13 15:23:18	480054	
out_17.bmp	2012-06-13 15:23:35	480054	
out_02.bmp	2012-06-13 15:23:35	480054	
out_22.bmp	2012-06-13 15:24:04	480054	
out_28.bmp	2012-06-13 15:23:54	480054	
out_21.bmp	2012-06-13 15:23:44	480054	
out_27.bmp	2012-06-13 15:23:44	480054	