Managing metadata

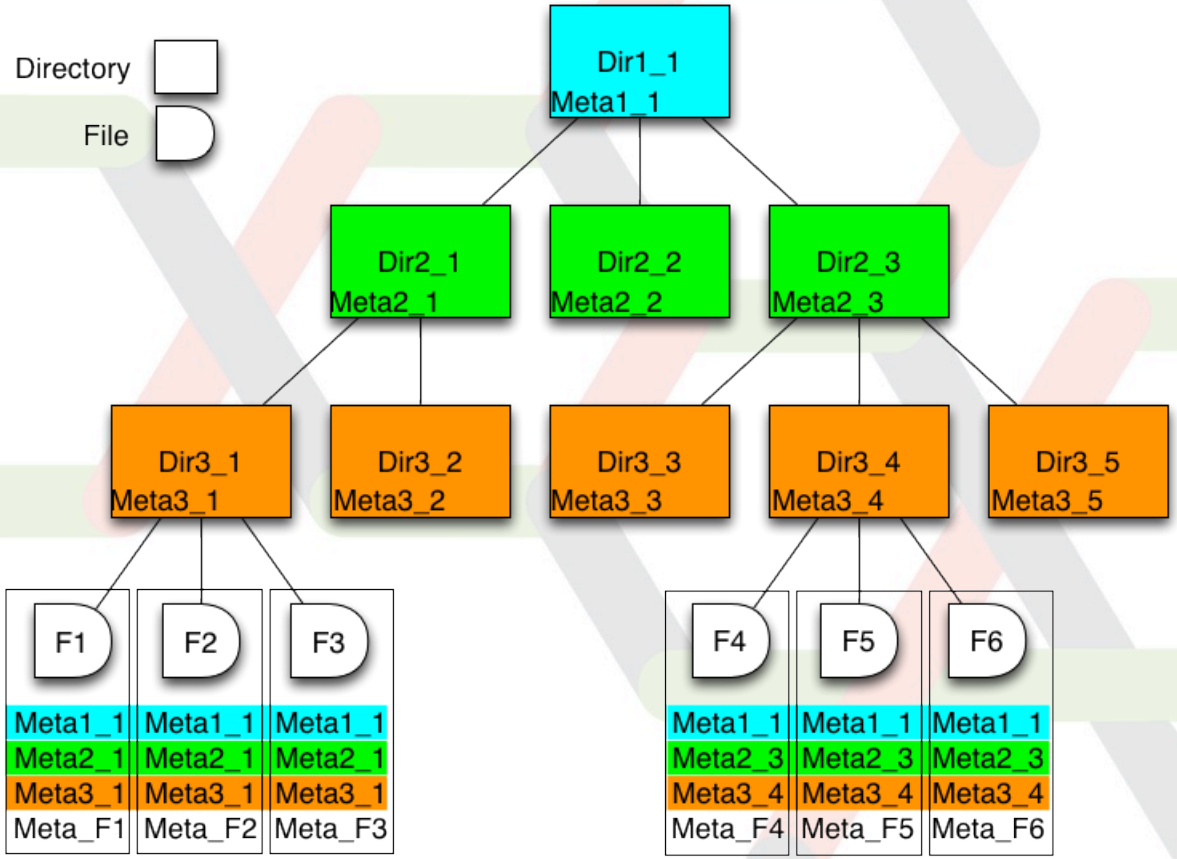
DIRAC Tutorial

FileCatalog Metadata

- User metadata can be associated to files and directories as key:value pairs
  - Value types: integer, float, string, date
- Some metadata can be declared globally as indices for further use in file selection / discovery
  - If not indexed, metadata can be considered as annotations

FileCatalog Metadata

- Files and subdirectories inherit metadata of their parents



Metaqueries

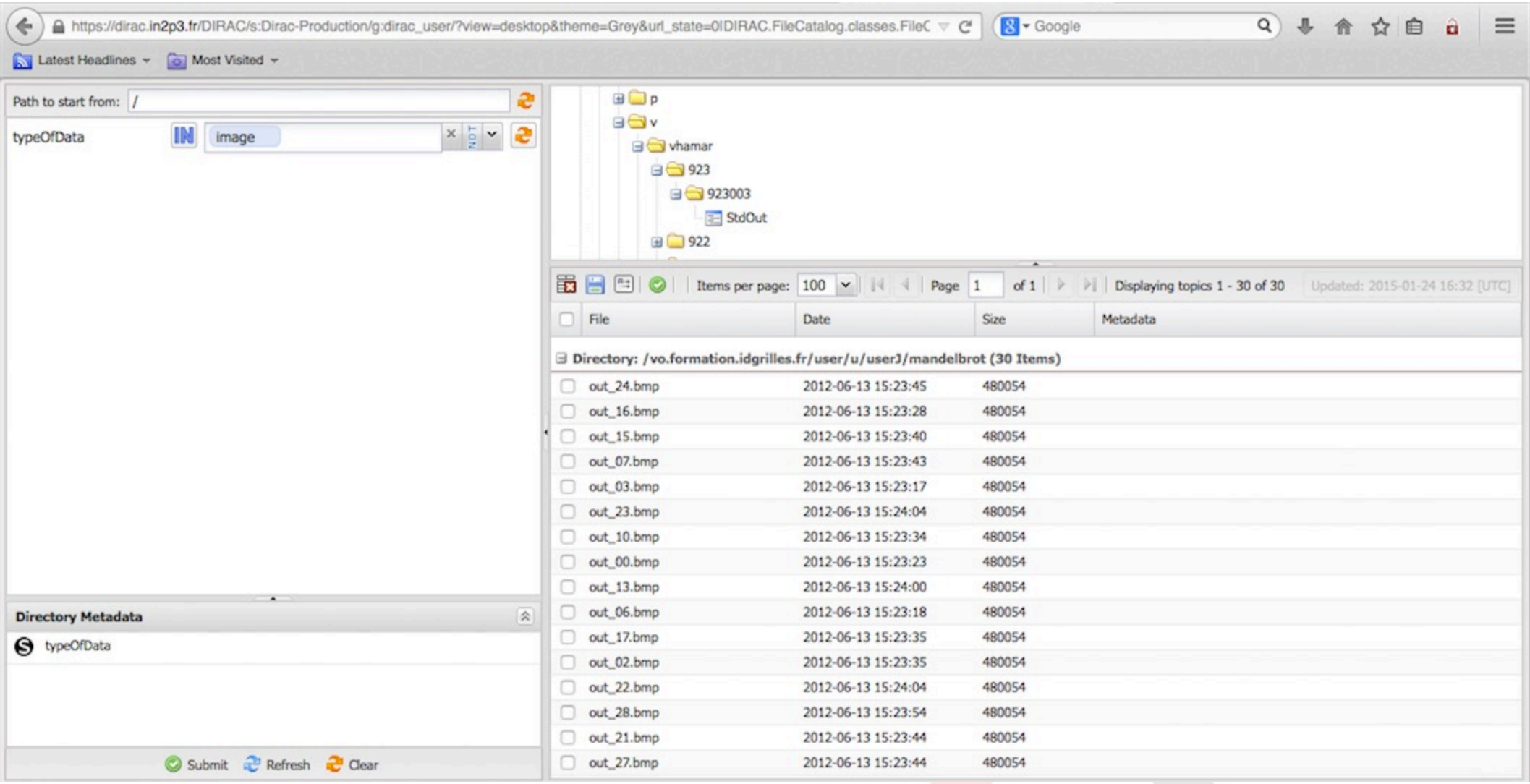
- Data selection with metadata queries
  - Example for DFC shell

```
find . Year=2010 Version=v1r0,v1r1 SE=CERN-disk
       ProductionTime>10-10-2013
```

- Web interfaces to make queries easier
  - Dynamically updated selectors as a function of already selected values
    - Available for directory metadata only

- COMDIRAC ***dfind*** command mimicking the unix ***find***

File Catalog Wep App



FileCatalog Metadata

- Functionality is similar to the AMGA gLite service or iRODS metadata
- Some extra features of DFC
  - Support the data provenance information
    - Ancestor<->descendent relationships
  - Support for efficient storage usage reports
    - Realtime
    - Necessary for the storage quota policies

Storage Usage

```
FC:/> size -l /lhcb/user/a/atsareg/1
directory: /lhcb/user/a/atsareg/1
Logical Size: 134,756,846 Files: 498 Directories: 500

    StorageElement  Size          Replicas
=============================================
  1 IN2P3-USER      20,254,050    75
  2 CNAF-USER       18,363,672    68
  3 RAL-USER        16,473,294    61
  4 CERN-USER       19,443,888    72
  5 GRIDKA-USER     21,064,212    78
  6 SARA-USER       20,254,050    75
  7 PIC-USER        18,903,780    70
---------------------------------------------
    Total           134,756,946   499
Query time 0.98 sec
```

- Report of storage usage for any directory
  - Whole community data
  - Per user data
  - “Logical” storage
    - LFNs, sum of the LFN sizes
  - “Physical” storage
    - Physical replicas, total volume per Storage Element

Support for datasets

- A relatively new feature developed for the BES III experiment in IHEP, Beijing
  - Datasets defined as a resulting list of files from a given metaquery
    - Particular case: all the files under a given directory
  - Dataset objects are stored in the same directory hierarchy as files
    - ACLs, ownership, show up in the output of ***ls*** command as for files
  - Datasets can be frozen in order not to change from one query to another
    - Can be refreshed by an explicit command, a quick check if changed since the last update
  - Datasets can be annotated
  - Operations on datasets
    - Replica lookup for all the files in a dataset
    - Total size, number of files report
    - More to come, e.g. replication, removal, etc

Tutorial

- Tutorial :
- https://github.com/DIRACGrid/COMDIRAC/wiki/Metadata-Management
- Add file and directory metadata
- Remove metadata
- Define metadata indices
- Search files depending on metadata values