Replicating Data from the Large Electron Positron (LEP) collider at CERN (Aleph Experiment)

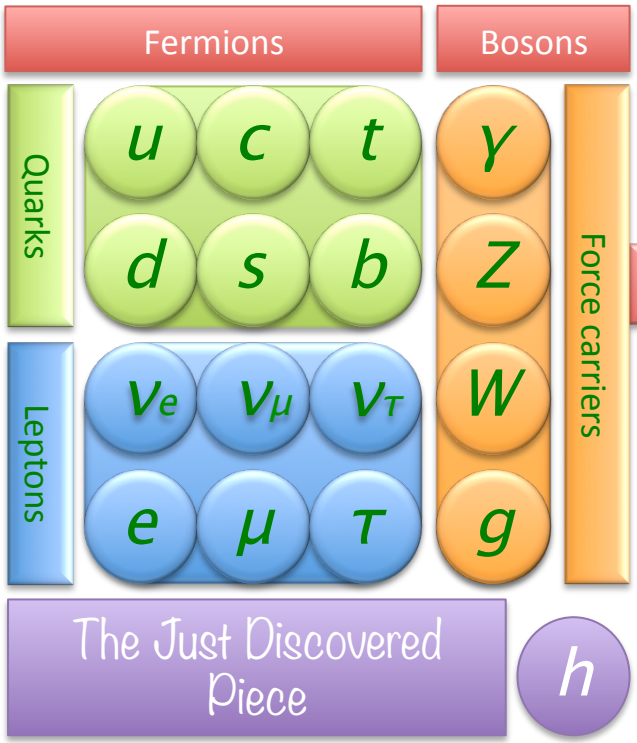
Marcello Maggi/INFN -Bari

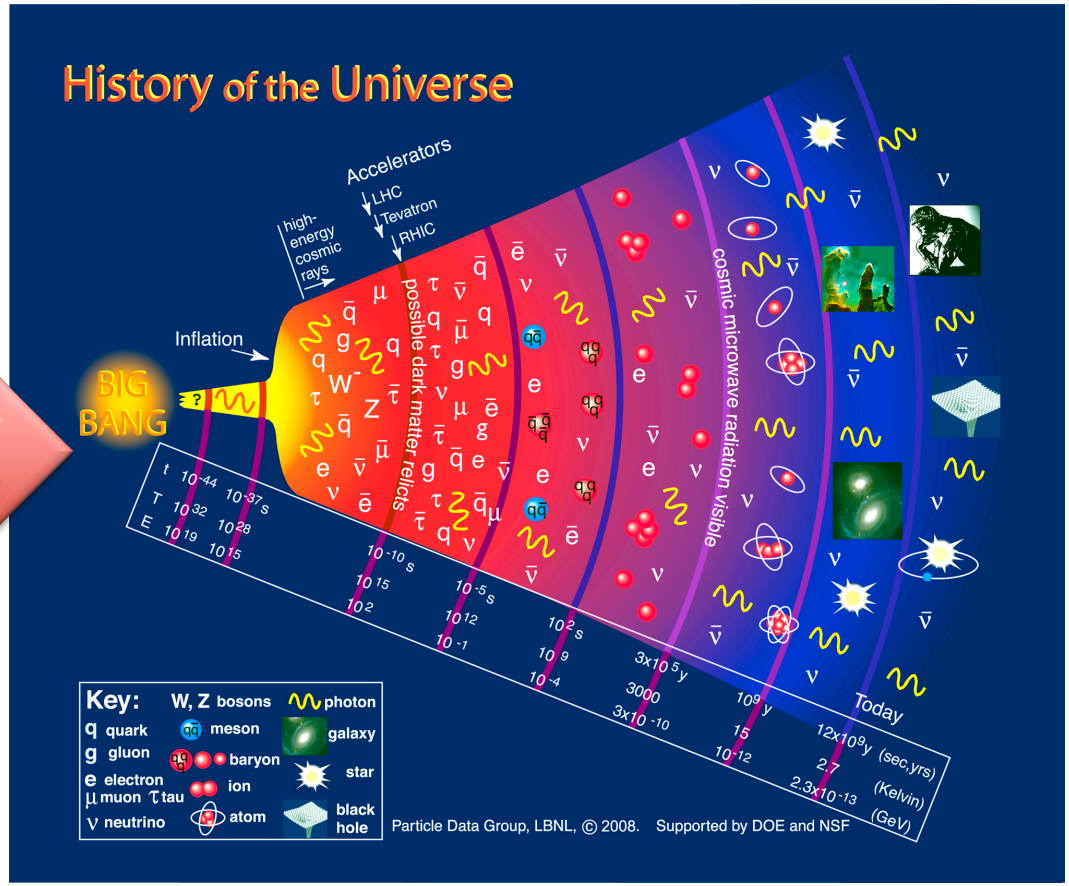
Tommaso Boccali/INFN-Pisa

Under the DPHEP umbrella

International Collaboration for Data Preservation and Long Term Analysis in High Energy Physics

The HEP Scientist

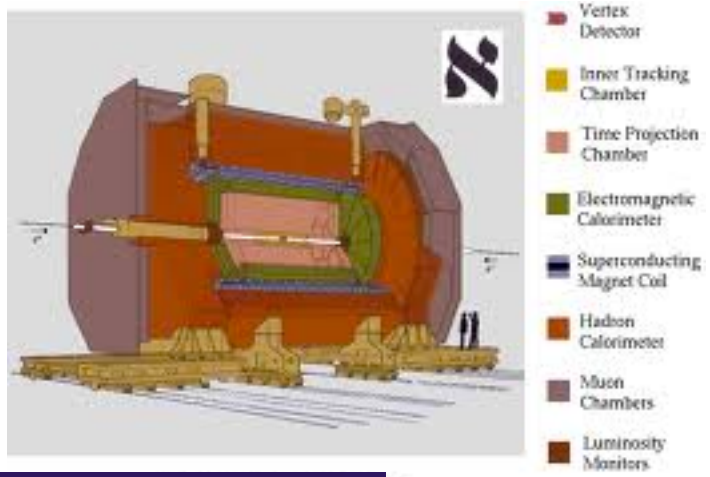


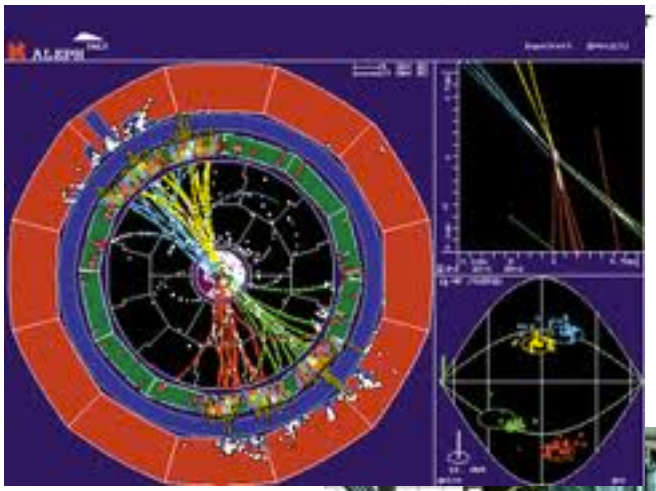


FROM MICROCOSM → Dark Matter, Matter/Anti-matter Asymmetry, Super Symmetric Particles → TO MACROCOSM

The use case

- The ALEPH Experiment took data from the CERN e^+e^- collider LEP from 1989 to 2000
- Collaboration of ~500 scientists from all over the world
- More than 300 papers were published by the ALEPH Collaboration





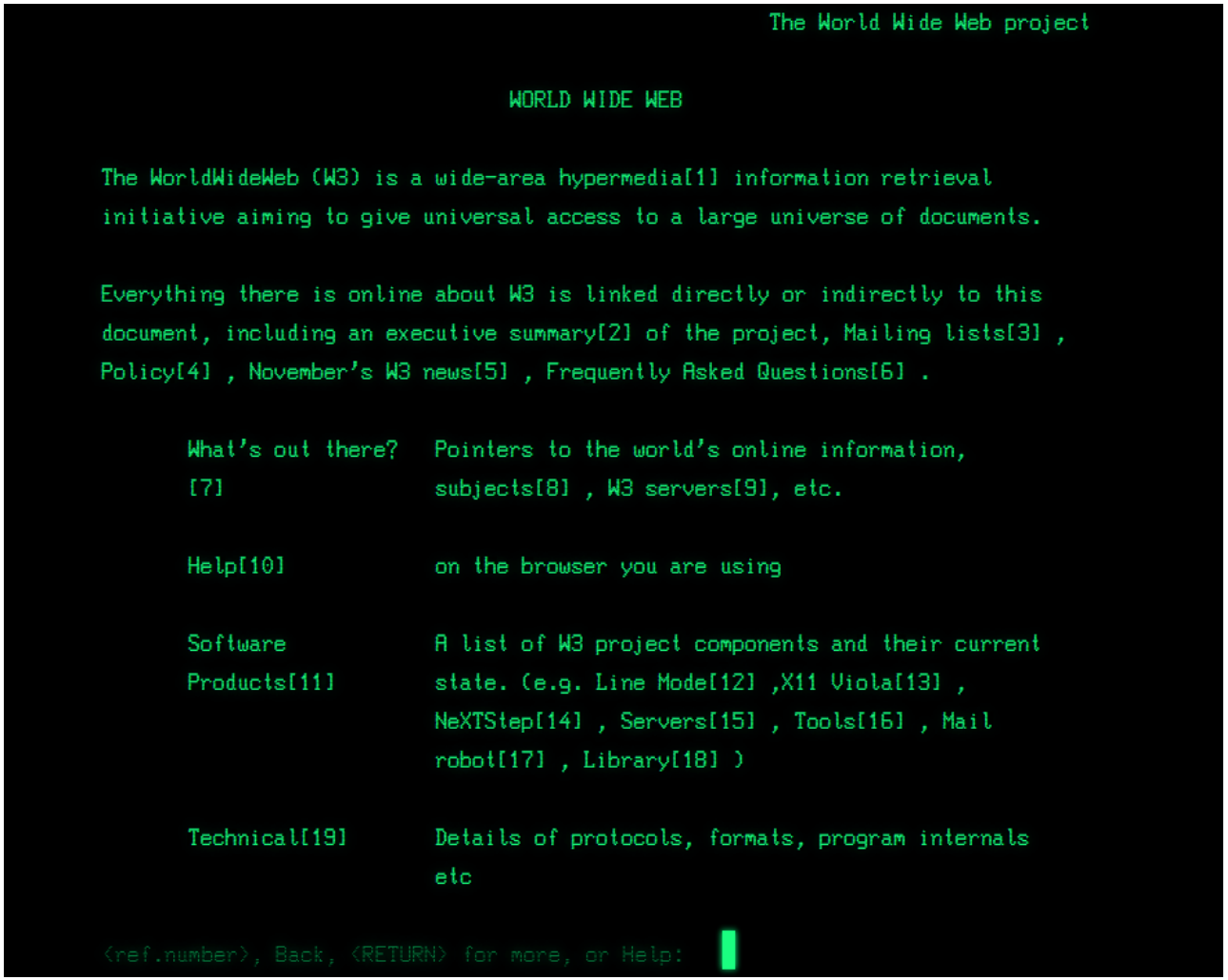
ALEPH data – still valuable

- While the Collaboration practically stopped a couple of years later, some late papers were published, and we still get request by Theoretical Physicists for additional checks / studies on ALEPH Data

- Current policy: any paper can use ALEPH data if among the author there is at least one former ALEPH Member (moving to CCO?)

Data Sharing &
Data Management
Fundamental Issue

Birth of Web @ CERN



Some Facts

Looking at Reconstructed samples
(the format closest to physics utilization)

ALEPH data consists in:

- 150k files, avg size 200 MB
- 30 TB
- Split between real collected events and Monte Carlo simulated events
- Processing times (analysis) on recent machines is such that 1 week is enough to process all the sample
  Splitting between machines helps!

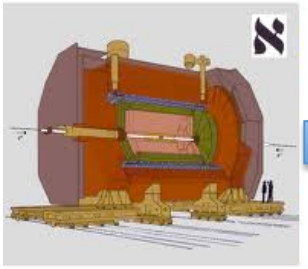
The Computing Environment

Last blessed environment (Blessed = blessed for physics) is Linux SL4

- GCC 3.4
- G77 3.4
- LIBC6
- All the sw ALEPH uses will have a CC licence

  We can recompile everything on our own

Data workflows



The ALEPH Detector

RAW DATA

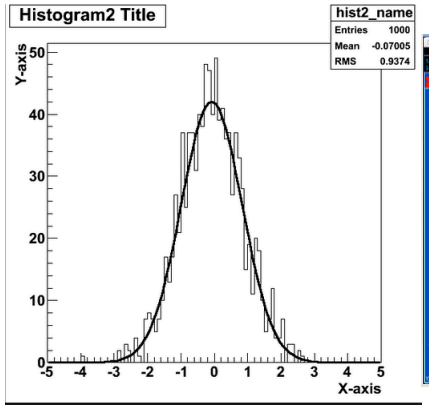
Reconstructed data

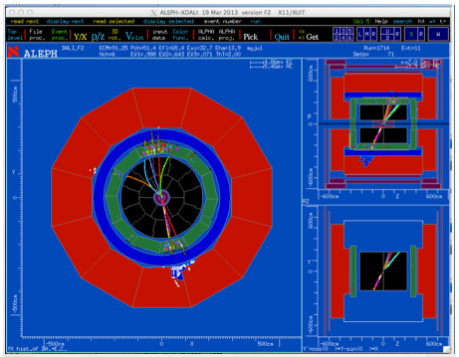
Fortran + HBOOK

User Code

Real DATA

Simulated Events





Kinematic Simulation

Paeticle/Matter Simulation

Reconstructed Simulated Events

Various tools

Geant3 based

“Julia”

Current Strategy

Computing Environment via VM approach

- Currently using uCERN-VM
- Provides batch ready VMs, interactive ready VMs, development ready VMs

Data to be served via POSIX to the executables

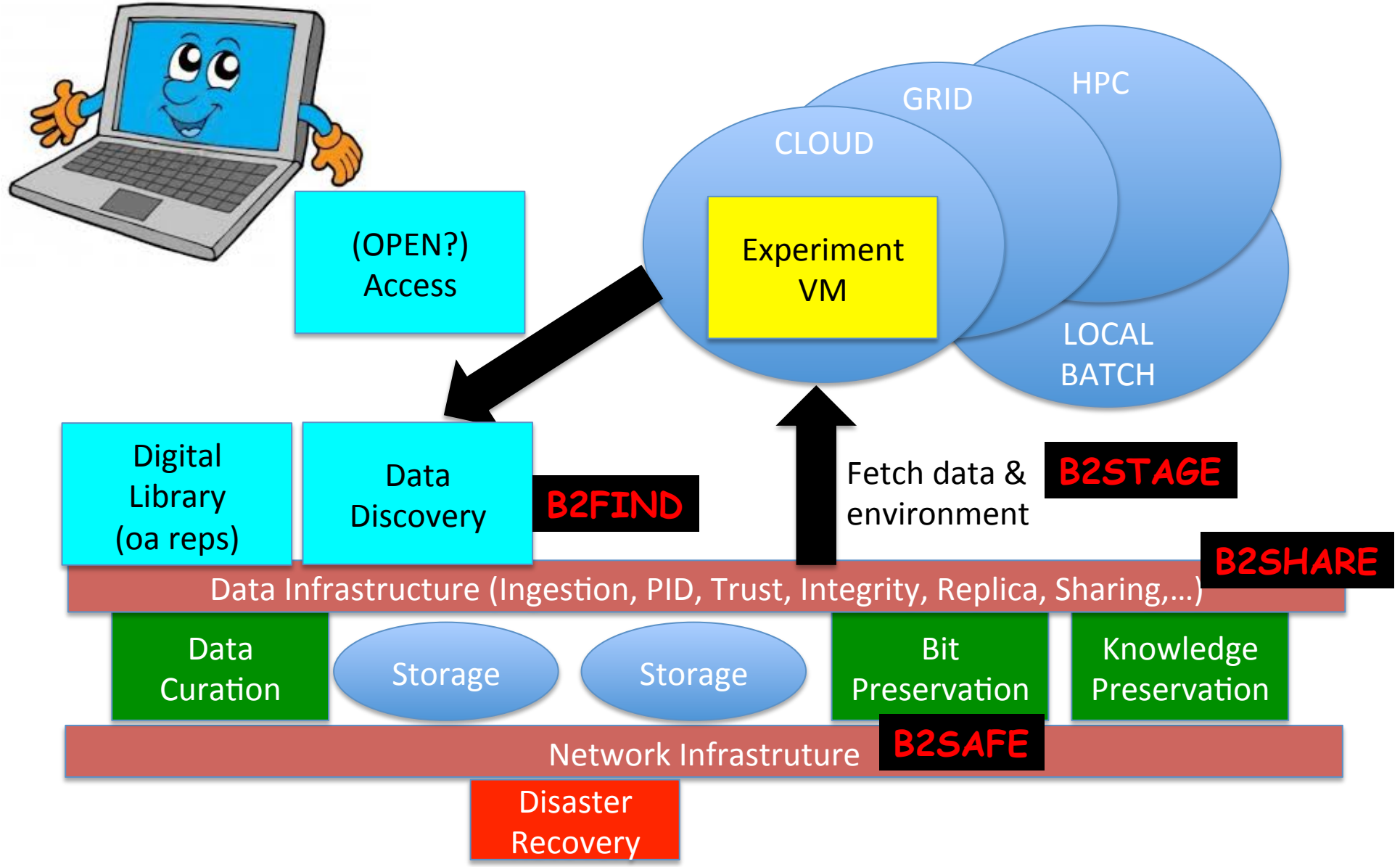
- Current approach (pre Eudat) was
  - Via WebDAV (Apache, Storm, ...)
  - Seen by the VM as FUSE/DavFS2 mounted POSIX Filesystem

What is missing?

- Metadata!
  - What is a file containing?
  - Where is a file available?
  - "Give me all the data taken at 203 GeV in Spring 1999"
  - "Give me all the simulations @ 203 GeV for hadronic W decays"

- We had a tool, SQL based
  - We still have the SQL dump
  - Tool only reachable via low level commands

The Picture


(OPEN?)
Access
HPC
GRID
CLOUD
Experiment
VM
LOCAL
BATCH
Digital
Library
(oa reps)
Data
Discovery
B2FIND
Fetch data &
environment
B2STAGE
B2SHARE
Data Infrastructure (Ingestion, PID, Trust, Integrity, Replica, Sharing,...)
Data
Curation
Storage
Storage
Bit
Preservation
Knowledge
Preservation
Network Infrastruture
B2SAFE
Disaster
Recovery

MetaData Ingestion
Complexity is the norm...

Welcome to INSPIRE, the High Energy Physics information system. Please direct questions, comments or concerns to feedback@inspirehep.net.

HEP :: HEPNAMES :: INSTITUTIONS :: CONFERENCES :: JOBS :: EXPERIMENTS :: JOURNALS :: AIUTO

Informazioni | Citazioni (0) | File

Data from the ALEPH collaboration, year 1999 (Real Data)

ALEPH Collaboration

Cite as: ALEPH collaboration (2013) EUDAT, http://doi.org/doi from EPIC server

Description: **** Debug of run content for this dataset

Run #	Nevents	Length (Mb)	LEP energy	Nb of Z0s
49217	4442	12.704	91.236	290
49218	648	1.955	91.234	44
49220	76	0.303	91.234	7
49222	3990	15.153	91.234	416
49223	6666	22.996	91.237	579
49224	5916	19.681	91.239	505
49235	6410	19.990	91.221	480
49240	11083	34.741	91.238	779
49241	10783	41.182	91.242	1054
49242	420	1.489	91.250	35
49243	11553	42.327	91.242	1076
49244	11697	46.579	91.242	1244
49245	11961	46.493	91.242	1251

Total length of information : 305.594 Mbytes

This dataset complements the following publication:
Update of the ALEPH non-strange spectral functions from hadronic τ decays

Record creato 2014-04-14, modificato l'ultima volta il 2014-04-14

Link to EUDAT

Export
BibTeX, EndNote, LaTeX(US), LaTeX(EU), Harvmac, MARC, MARCXML, NLM, DC

Eudat?

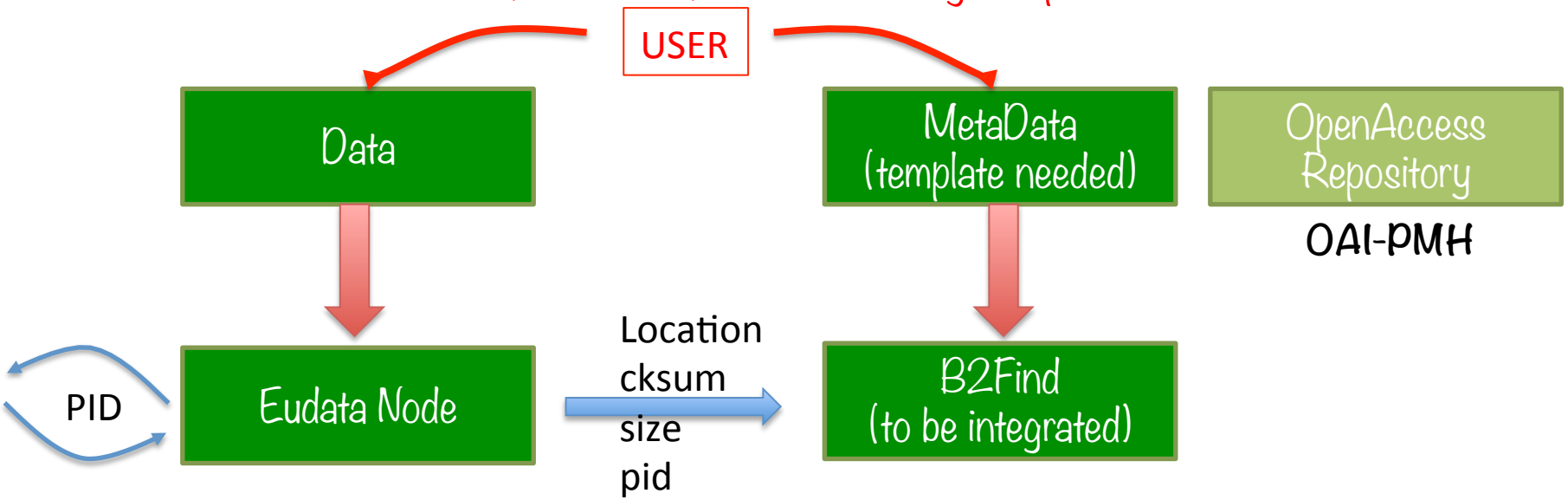
Plan is "simple"

- Move all the files WebDAV -> Eudat node
  - Already started, to CINECA-Eudat node
- Use B2FIND to re-implement a working metadata catalog (SQL to B2FIND??)
- Using other B2* tools to (we need to study here)
  - Prestage files on the VM before execution? (B2STAGE)
  - Access via streaming? (B2STREAM?)
- Using B2* to save outputs in a safe place, with correct metadata (B2SHARE)

Few Starting Feedbacks

150,000 files... Missing bulk "B2Ingestion"

B2SHARE, B2SAFE,B2STAGE do "just" part of it



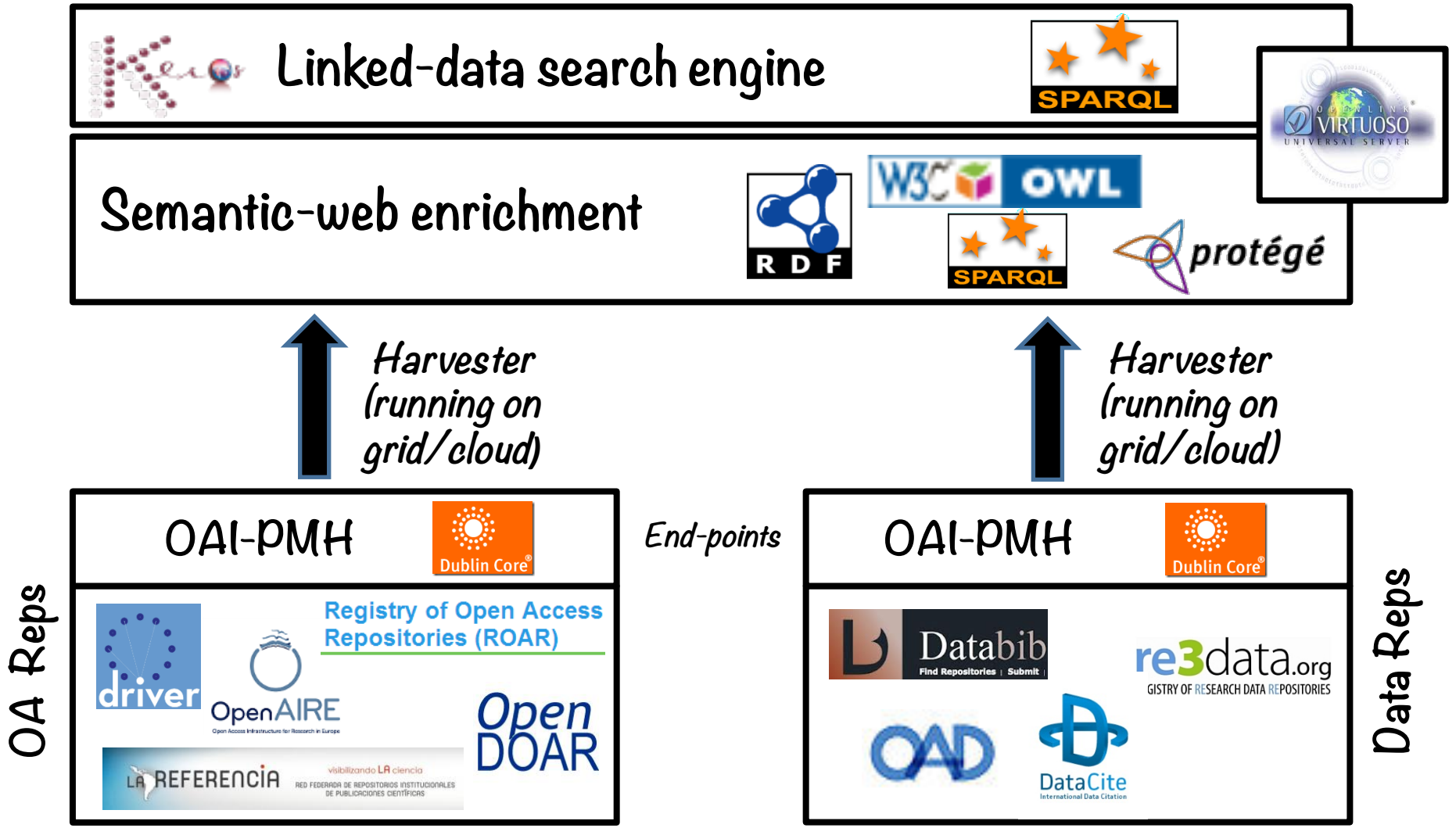
How it is now:

1) Provide identity 2) gridftp 3) copyback pids

4) build metadata 5) feed OA reps 6) give oai-pmh link

Roberto Barbera/INFN e Università, Catania

Data & Knowledge Infrastructure



Knowledge Base

Not only data

- ALEPH Virtual Environment

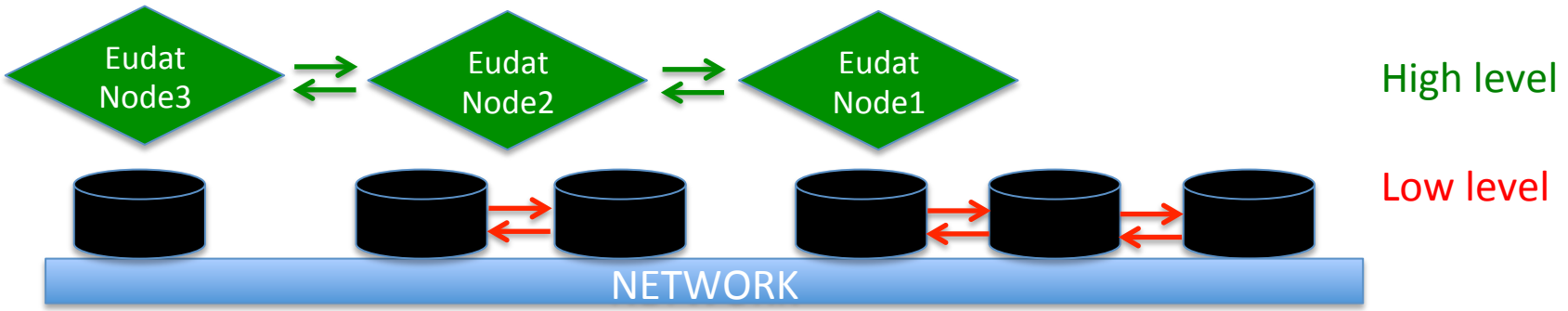
Link to Digital Library (Open Access Reps)

- File Catalogues
- Documentation
- Instructions

HEP status

- ALEPH is now a small experiment
- KLOE has 1 PB
- CDF has 10 PB
- LHC has > 100 PB

Data Nodes can federate, specially useful for sites where computing is available



The (Big) DATA today

10^7 "sensors" produce 5 PByte/sec
Complexity reduced by a Data Model
Analytics in real time filters to 0.1–1 Gbyte/sec (Trigger)

Data + Replica move with a Data Management Policy
Analytics produce "Publication Data" that are Shared
Finally the Publications

Re-use of Data relies on Data Archive Management Plans