Long Term Data Preservation

Frank Berghaus

On Behalf of the DPHEP Collaboration

Objectives

- Preserve data, software, and know-how in the collaborations:
  - Data/Bit Preservation (MoU between CERN and tier1's)
  - **Analysis Preservation**
    - **Preserve software evolution alongside data**
  - Timescale: "Forever" 30+ years past experiment life
- Share data, software, and know-how:
  - Larger Scientific Community
  - Education and Outreach
- CernVM is well placed:
  - Many experiments (LHC and others) already use CernVM and CVMFS
  - Virtualization is ubiquitous, and probably won't go away

Data Preservation & Sharing

- LHC experiments defined strategy and scope defined in these policy documents
  - Many overlapping requirements and tasks!
- Example experiments preserving data:

Experiment	Approach
BaBar	Virtual machines & Infrastructure servers
HERMES	Web/wiki, Logbook, Mailing lists, dcache, AFS, BIRD batch, GRID
ALEPH	Running code in a SL6 VM, Open to CernVM & CVMFS
Belle	Reformatting data for BelleII software
CDF	Plans to preserve all data & software, R&D stage

Analysis Preservation

- Goal: Reproducibility for the collaboration
- Capture physics analysis code (Snapshot?)
  - Libraries and compiler -> depend on hardware
    - Virtualization easier to port than individual software
  - Analysis metadata
    - Experiment conditions
    - Software provenance
  - Input Data

Analysis Preservation: Data

- Input data
  - RAW data and Reconstruction code
    - Software provenance & full database (alignment, conditions, etc.)
  - **Or** capture input data with analysis per analysis
    - Limits scientific reach
    - Data volume may not be feasible to capture
- Capture analysis and production software environment

Capture Approaches

- Store static virtual machine with analysis
  - Capture all code, database, and environment in single large (many GB) VM image
- Use a set of Docker containers for each analysis
  - Create system containers each hosting a part of the necessary services
  - Idea: Using docker to capture analysis infrastructure and using CernVM as consistent host?
- Contextualize CernVM from analysis metadata
  - Use CVMFS to provide operating system, compilers, experiment software & databases, and external libraries for physics analysis code

Analysis Capture Framework

- Invenio as mechanism to capture publications and metadata

  invenio-software.org

  - Contextualize or create service orchestration from metadata
- Demonstration portal for analysis and metadata capture is

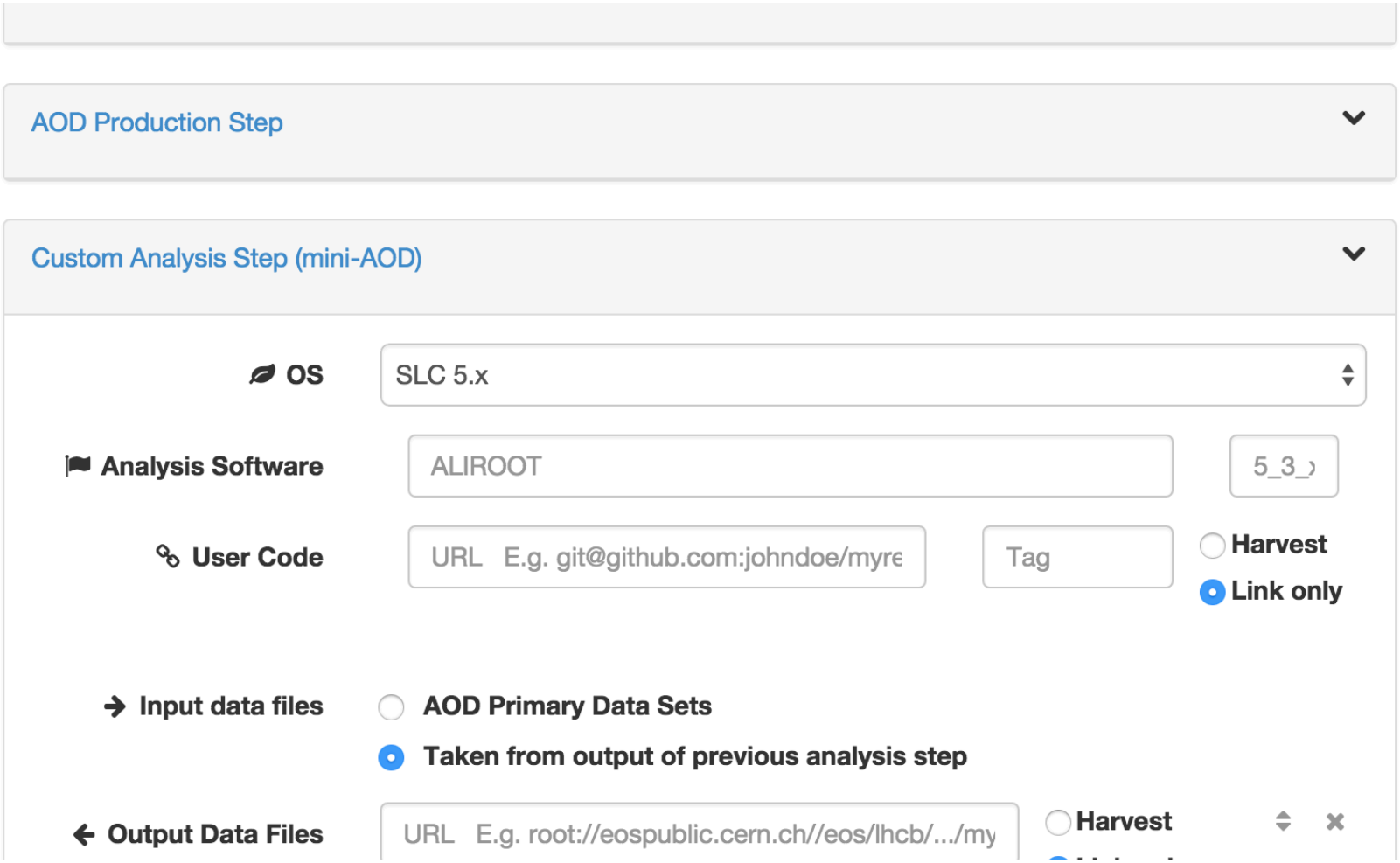
  analysis-preservation.cern.ch

Analysis Preservation Demo Portal

Acess to all submitted data will be restricted to the ALICE collaboration only.

- Basic Information
- Physics Information
- AOD Production Step
- Custom Analysis Step (mini-AOD)
- End-user analysis
- Internal Documentation

Analysis Preservation Demo Portal

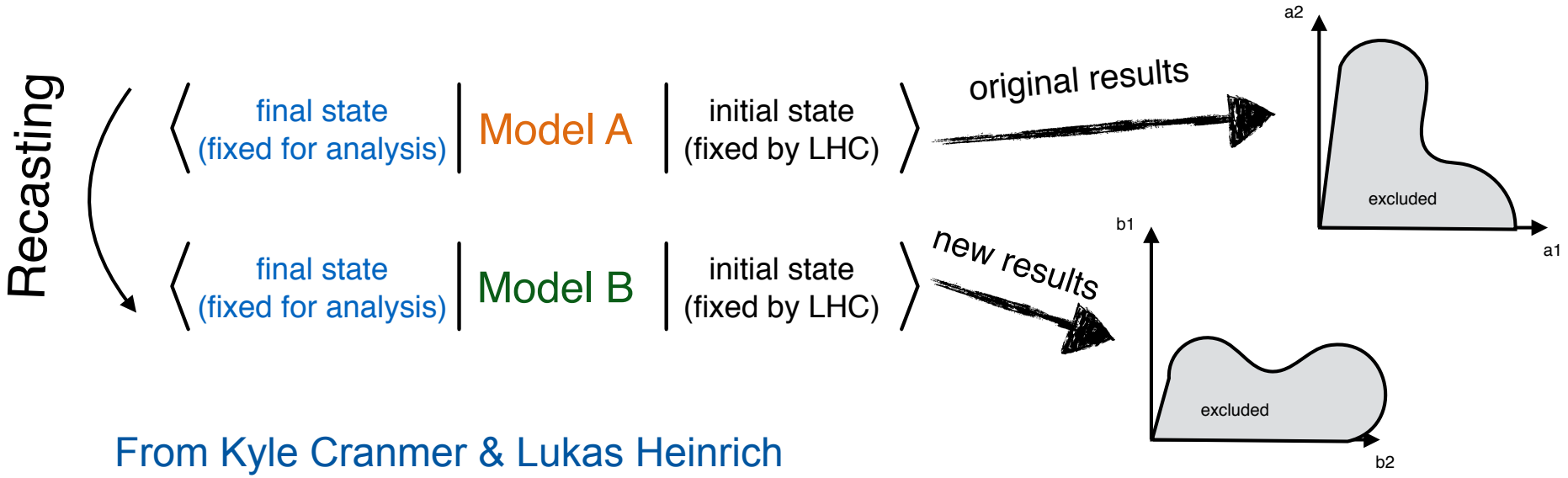


Metadata capture

- Goal: script to capture analysis environment information
  - Upload information to the portal
  - Retrieve software provenance
    - Parse input data through existing experiment provenance systems (AMI, etc.)
    - Could this be used with cvmfs tagging?
  - Analysis preservation portal to allow additions, modifications, review, and archival
- What helper scripts/API already exist?

Running Captured Analyses

- Start with RECAST
  - Capture analysis for Model A with final state
  - Model B has same final state
  - Reinterpret captured analysis under new model



From Kyle Cranmer & Lukas Heinrich

RECAST

- Frontend accepts requests for processing:
  - recast.perimeterinstitute.ca
- Collaboration approves requests for batch processing at the control center:
  - recast-demo.cern.ch
- Backend processing is done on CernVM running CERN OpenStack
  - Software requirements (Repository suggestions?):
    - Rivet - analysis infrastructure
    - Fast simulation - FastSim, ATOM, etc.
    - Full experiment simulation
  - Scalable processing backend
    - Condor/cloud scheduler – proven to work with CernVM
    - OpenStack HEAT template?

Ideas from yesterday

Open Access: Science

- Reinterpreting existing analysis for new model
  - Easy to use
  - Needs interface & resources
  - ATLAS: Developing RECAST
- Access to experiment software, data, and simulation
  - Introduced in Kati Lassila-Perini's talk on Thursday
  - Follow "Research" at: http://opendata.cern.ch/
    - Ioannes' WebAPI would be amazing here!
  - CMS: 50% (~27TB) of 2010 data released
  - ALICE: Will release 10TB of 2010 data this year

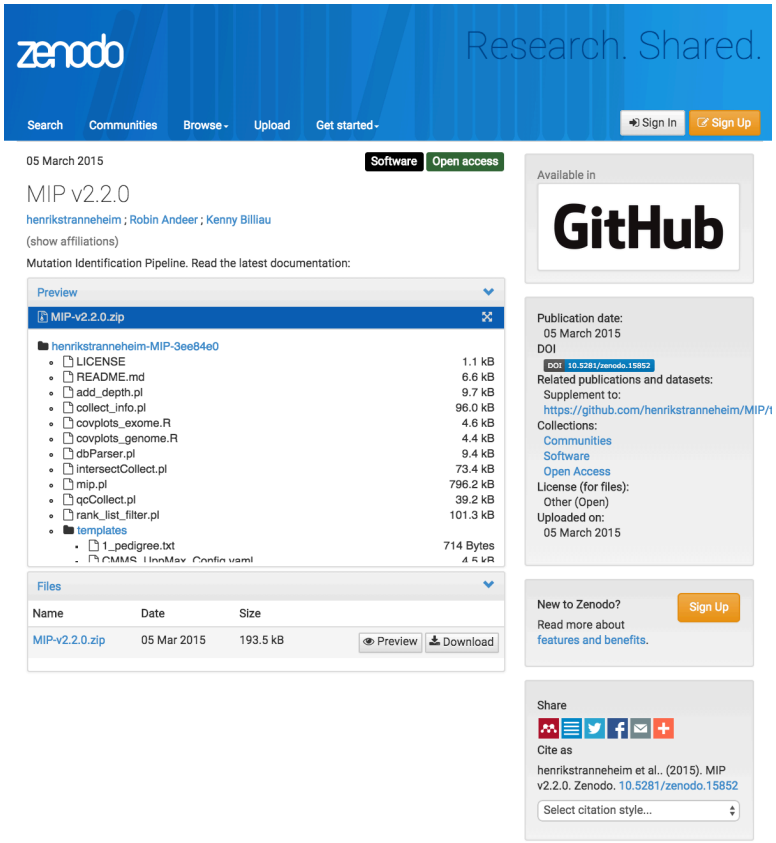
Summary/Questions

- CernVM is a great candidate for preservation:
  - LHC experiments and some others leverage CernVM and CVMFS already
  - What about supporting other disciplines (i.e. non SL distributions)?
- Could we distribute small (analysis-level) databases via cvmfs?
  - What about large databases needed for production?
  - Where would cvmfs versioning be useful?

Backup

Aside - Zenodo: Fringe science

- Stores publication with data and code
- Often generic code using open source tools, e.g. R, SciPy
- CernVM WebApp could be useful to create test analysis environment?



Data Preservation outside HEP

- Examples from outside HEP:
  - Space & Astronomy: NVO, EURO-VO
  - Earth & Ocean: PANGEA,
  - Life Sciences: ELIXIR
  - Minimize software dependence by using ubiquitous and well documented standard formats
    - No current standard data format for HEP
    - HEP Experiments, simulation, and analysis require complex tools

CVMFS For Preservation

- Pros:
  - Many experiments already use cvmfs
  - CERN and Tier1's agreed to maintain infrastructure (MoU)
- Cons:
  - Requires infrastructure
  - Experiment must use CVMFS
- Use beyond LHC experiments
  - Is it reasonable to expect CVMFS adoption?
  - Ease of using non-SL, non-RPM linuxes?

Existing Approaches in HEP

	BaBar	H1	ZEUS	HERMES	Belle	BESIII	CDF	D0
End of DAQ	07.04.08	30.06.07	30.06.07	30.06.07	30.06.10	2017	30.09.11	30.09.11
OS	SL 3/5 RHEL 3/5	SL 5	SL 5	SL 3/5	SL 5 RHEL 5	SL5	SL 5/6	SL5
Languages	C++ Python Java	C C++ Fortran Python	C++	C C++ Fortran Python	C C++ Fortran	C++	C C++ Python	C++
Simulation	GEANT 4	GEANT 3	GEANT 3	GEANT 3	GEANT 3	GEANT 4	GEANT 3	GEANT 3
External Dep's	ACE CERNLIB CLHEP CMLOG Flex GNU Bison MySQL Oracle ROOT TCL XRootD	CERNLIB FastJet NeuroBays Oracle ROOT	ROOT	ADAMO CERNLIB ROOT	Boost CERNLIB CLHEP NeuroBays PostgresQL ROOT	CASTPR CERNLIB CLHEP HepMC ROOT	CERNLIB NeuroBayes Oracle ROOT	Oracle ROOT

From: DPHEP-2012-001

Existing Approaches at CERN

	ALICE	ATLAS	CMS	LHCb	ALEPH	DELPHI	L3	OPAL
End of DAQ	~2030	~2030	~2030	~2030	02.11.00	02.11.00	02.11.00	02.11.00
OS	SL 6?	SL 6	SL 5/6	SL 6?	?	?	?	?
Languages	C++ Python	C++ Fortran Python	C++ Python	C++ Python	?	?	?	?
External Dep's	ROOT ?	ROOT ?	ROOT ?	ROOT ?	CERNLIB ?	?	?	?

From: DPHEP-2012-001

Open Data: Open Access

- Access to experiment software and software
  - See “Research” at: http://opendata.cern.ch/
    - Note: Ioannes’ WebAPI would be amazing here!
  - HEP data requires custom, and complex code
    - For each experiment
  - Code usually not portable, requires:
    - Specific libraries
    - Compiler version
    - Operating system & architecture
- Virtualization is ubiquitous and provides:
  - Libraries, compilers, and software tools

Existing Approaches in HEP

Experiment	Approach
BaBar	Virtual machines & Infrastructure servers
H1	People doing analysis, Data stored
ZEUS	People doing analysis, Data stored
HERMES	Web/wiki, Logbook, Mailing lists, dcache, AFS, BIRD batch, GRID
Belle	Developing BelleII code to be backwards compatible
BESIII	Thinking about preservation
CDF	Plans to preserve all data & software, R&D stage
D0	Under discussion, fraction of a FTE working on it

From: DPHEP-2012-001

- BaBar showed that virtualization works for preservation.

Existing Approaches at CERN

Experiment	Status
ALICE	Interested in CernVM & cvmfs
ATLAS	Investigating docker, forward porting data, open to CernVM
CMS	Interested in CernVM & cvmfs, forward porting data
LHCb	Interested in CernVM & cvmfs
ALEPH	Porting code to modern OS & compilers. Building a VM, Open to CernVM & CVMFS
DELPHI	?
L3	?
OPAL	Porting code to modern OS & compilers. Building a VM, Open to CernVM & CVMFS

- Even forward porting data requires validation against old release.