HEPiX bit-preservation WG

Dmitry Ozerov/DESY
Germán Cancio/CERN

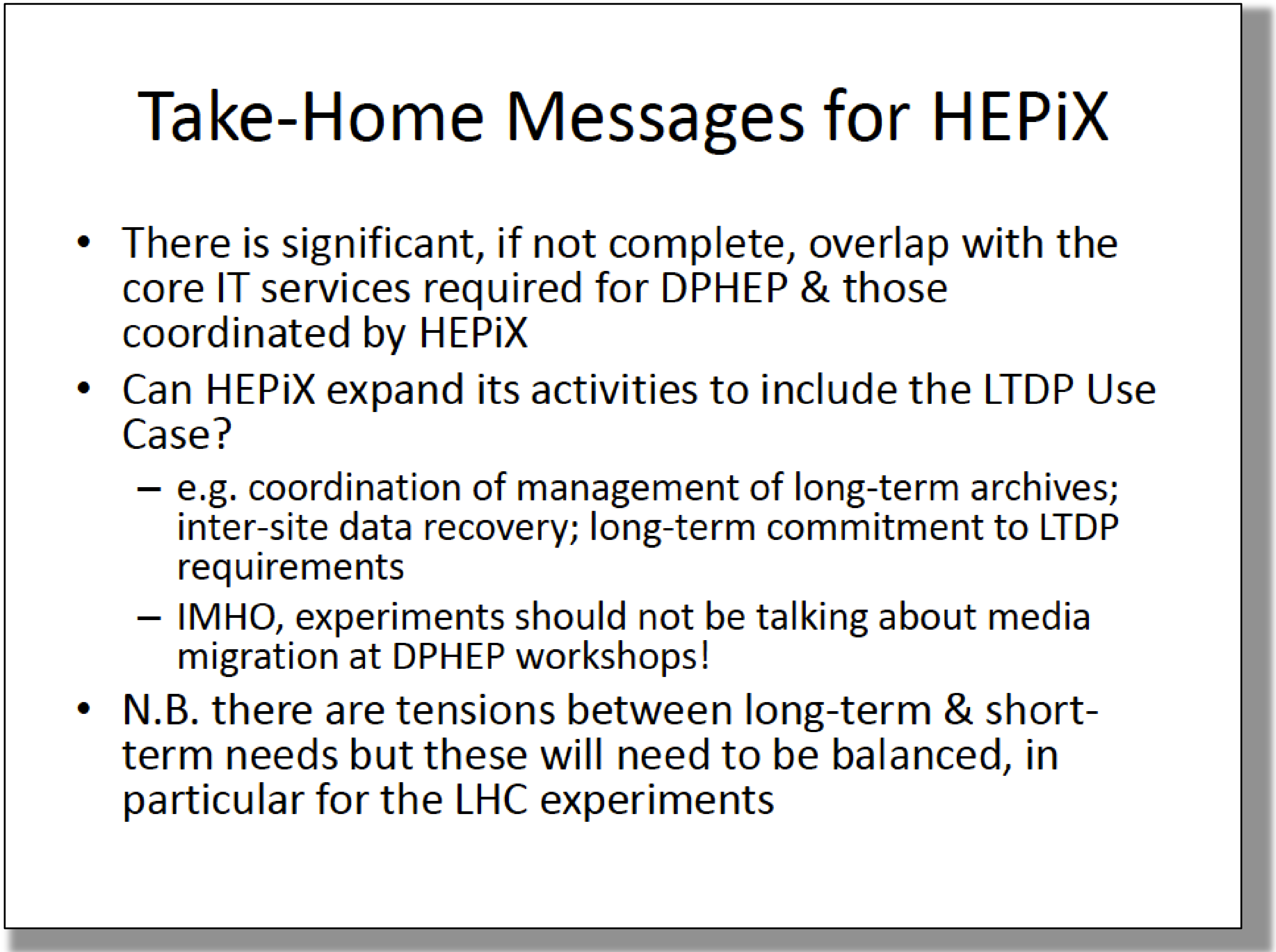
HEPiX Fall 2013, Ann Arbor

Agenda

- Bit-preservation WG: background, mandate
- Large archive sites survey
  - – main findings
- Conclusions & next steps

Bit-preservation WG: Background

- Request from DPHEP @ HEPiX spring 2013:



Take-Home Messages for HEPiX

- There is significant, if not complete, overlap with the core IT services required for DPHEP & those coordinated by HEPiX
- Can HEPiX expand its activities to include the LTDP Use Case?
  - e.g. coordination of management of long-term archives; inter-site data recovery; long-term commitment to LTDP requirements
  - IMHO, experiments should not be talking about media migration at DPHEP workshops!
- N.B. there are tensions between long-term & short-term needs but these will need to be balanced, in particular for the LHC experiments

Bit-preservation WG: Mandate

- The goal of the HEPiX Bit Preservation Working Group is to share ideas, practices and experience on bit stream preservation activities across sites providing long-term and large-scale archive services. Different aspects should be covered like: technology used for long-term archiving, definition of reliability, mitigation of data loss risks, monitoring/verification of the archive contents, procedures for recovering unavailable and/or lost data, procedures for archive migration to new-generation technology.
- The Working Group responds to a request by the DPHEP collaboration for advice on technical matters of bit preservation.
- The Working Group will produce a survey on existing practices across HEPiX and WLCG sites responsible for large-scale long-term archiving. The collaboration should ideally be extended to other large-scale archive sites from other research fields outside HEP.
- Based on best practices and development in storage preservation activities, the Working Group will provide recommendations for sustainable archival storage of HEP data across multiple sites and different technologies.

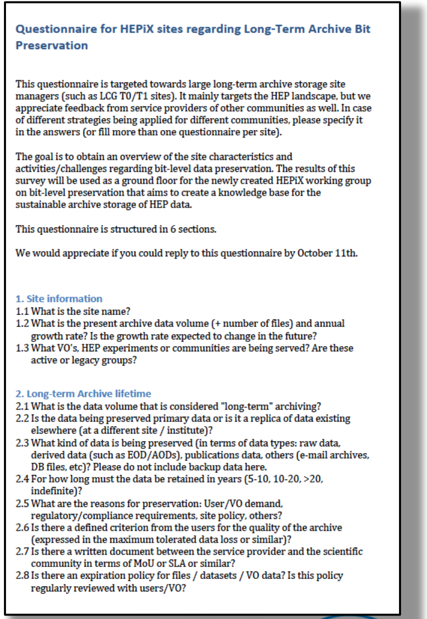
Bit-preservation WG: Mandate

- Collecting and sharing knowledge on bit preservation across HEP (and beyond)
- Provide technical advise to DPHEP
- Recommendations for sustainable (distributed) archival storage in HEP

w3.hepix.org/bit-preservation

Large archive sites survey

- Targeted towards large long-term sites (such as LCG T0/T1 sites)
- Goal is to obtain an overview of the site characteristics and activities/challenges regarding bit-level data preservation.
- 40 questions structured in 6 sections
- Survey sent to 21 sites, received 18 answers: THANKS
  - 2 sites had no input to provide ("young" sites, no archiving experience)
  - All "large" HEPiX/WLCG Tier-1 sites provided input



Questionnaire for HEPiX sites regarding Long-Term Archive Bit Preservation

This questionnaire is targeted towards large long-term archive storage site managers (such as LCG T0/T1 sites). It mainly targets the HEP landscape, but we appreciate feedback from service providers of other communities as well. In case of different strategies being applied for different communities, please specify it in the answers (or fill more than one questionnaire per site).

The goal is to obtain an overview of the site characteristics and activities/challenges regarding bit-level data preservation. The results of this survey will be used as a ground floor for the newly created HEPiX working group on bit-level preservation that aims to create a knowledge base for the sustainable archive storage of HEP data.

This questionnaire is structured in 6 sections.

We would appreciate if you could reply to this questionnaire by October 11th.

1. Site information

1.1 What is the site name?

1.2 What is the present archive data volume (+ number of files) and annual growth rate? Is the growth rate expected to change in the future?

1.3 What VO's, HEP experiments or communities are being served? Are these active or legacy groups?

2. Long-term Archive lifetime

2.1 What is the data volume that is considered "long-term" archiving?

2.2 Is the data being preserved primary data or is it a replica of data existing elsewhere (at a different site / institute)?

2.3 What kind of data is being preserved (in terms of data types: raw data, derived data (such as EOD/AODs), publications data, others (e-mail archives, DB files, etc)? Please do not include backup data here.

2.4 For how long must the data be retained in years (5-10, 10-20, >20, indefinite)?

2.5 What are the reasons for preservation: User/VO demand, regulatory/compliance requirements, site policy, others?

2.6 Is there a defined criterion from the users for the quality of the archive (expressed in the maximum tolerated data loss or similar)?

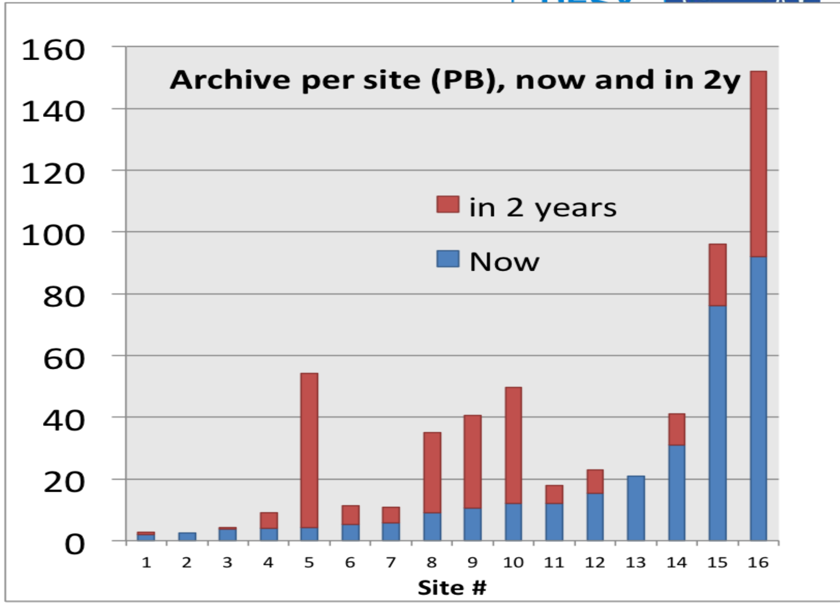
2.7 Is there a written document between the service provider and the scientific community in terms of MoU or SLA or similar?

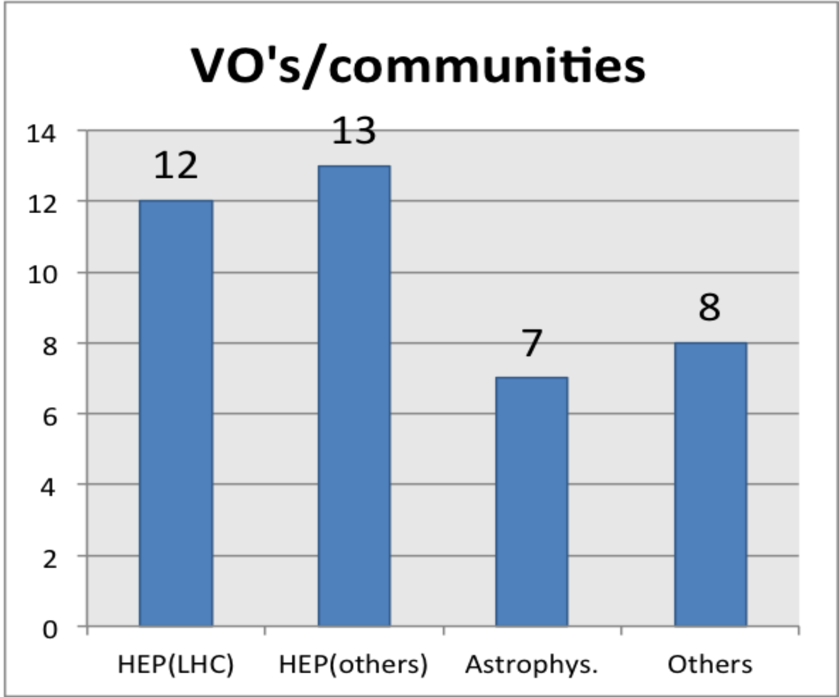
2.8 Is there an expiration policy for files / datasets / VO data? Is this policy regularly reviewed with users/VO?

Survey: Site information

- Per-site archive volume: 2-92PB
  - Total archive space = 306 PB
  - Future yearly growth 0% - 10x
  - > 0.57 EB (+86%) expected end 2015!

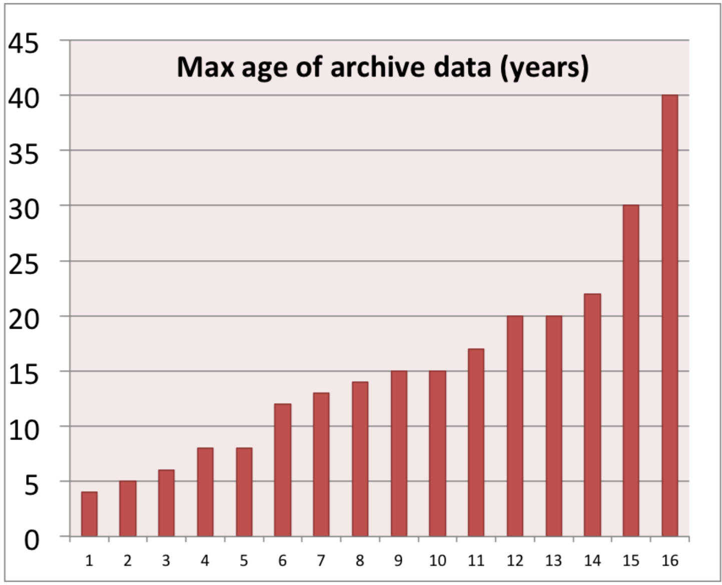
- Served VO's/communities:
  - 2 sites serving only LHC
  - One site serving a single community
  - 8 sites serving non-running HEP experiments



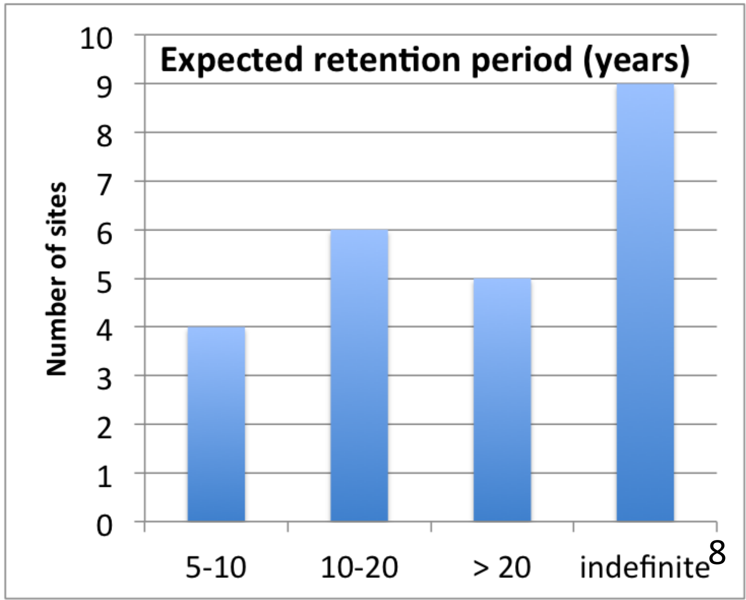


Survey: Archive lifetime(1)

- Vast majority of sites see **all** archived data as “long-term” (297PB out of 306PB)
- ALL sites store custodial data (no replica elsewhere). LHC data – mostly replicated; non-LHC: mostly not!
- Current archived data dating back 4..40 years

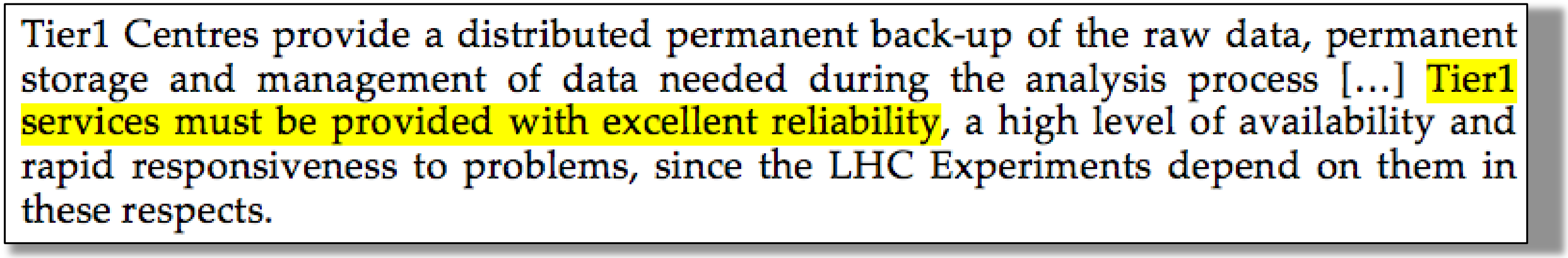


- Expected retention period: 5y.. indefinite (≠ infinite!)
- Only 2 sites have a data expiration policy
  - 4 sites admit preserving data “for no good reason” (unclear utility / ownership)
- Only 4 sites confirm that budget is secured for the complete preservation lifetime!



Survey: Archive lifetime(2)

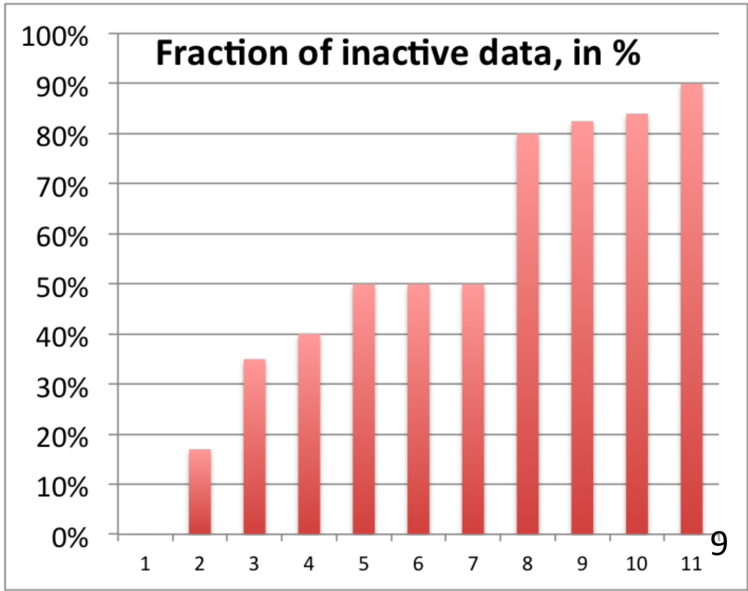
- Only 4 sites have an SLA or QoS user agreement wrt data loss. 8 have signed the WLCG MoU…



Tier1 Centres provide a distributed permanent back-up of the raw data, permanent storage and management of data needed during the analysis process [...] Tier1 services must be provided with excellent reliability, a high level of availability and rapid responsiveness to problems, since the LHC Experiments depend on them in these respects.

- Most sites quoting “no data loss” as user expectancy(!)

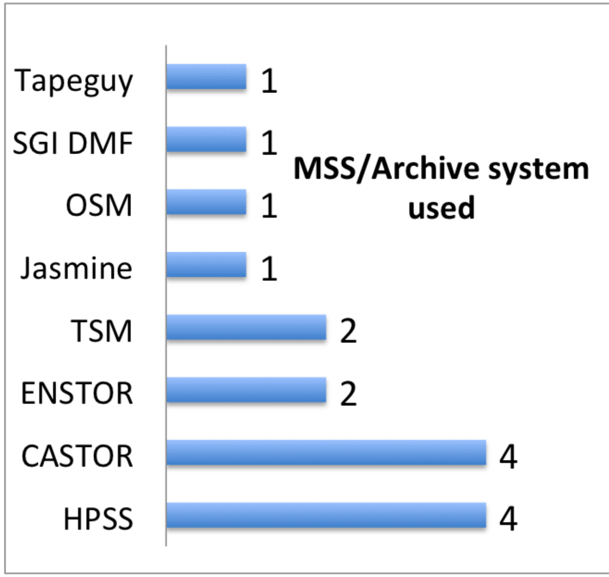
- Large fractions of “cold” data (average: 53%)
  - “cold” == not accessed > 12 months



Fraction of inactive data, in %

Survey: Archive Storage / Access

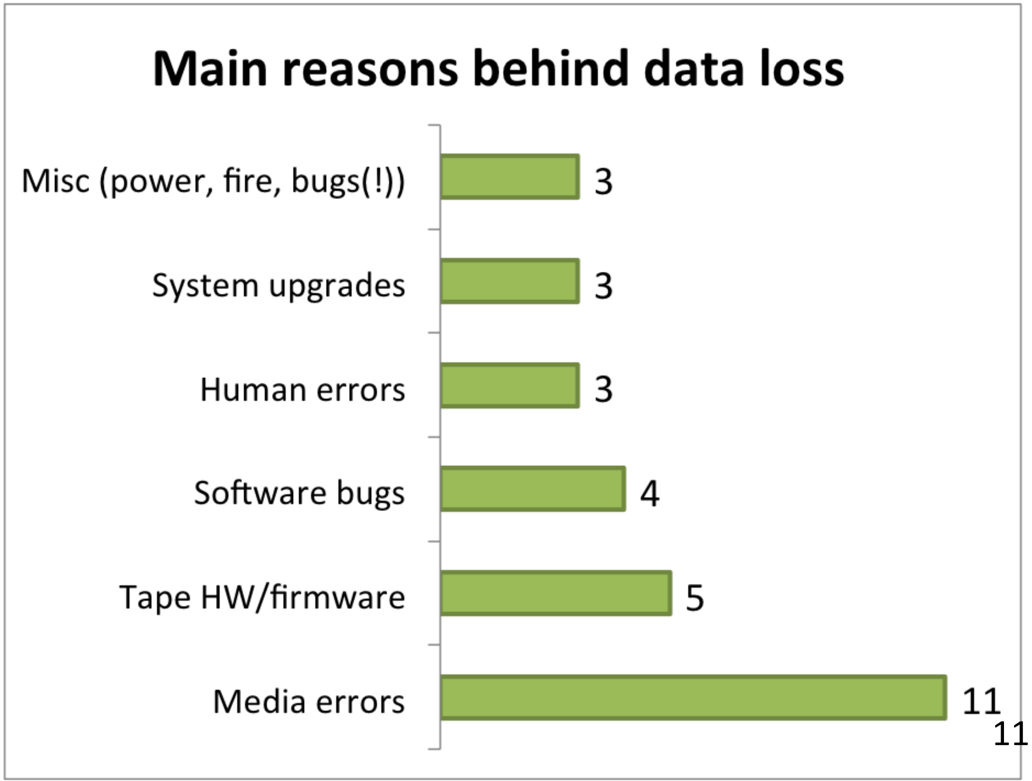
- All but one sites: No dedicated archiving solution, main MSS == main archive
  - Not an intentional choice, but became so de-facto
- 8 different systems (commercial/OSS)
  - 3 single-instance tools
  - Staffing ~ >=1 FTE (more for larger sites)



- On all sites data is available online or “nearline” (robotic libraries)
- Data is accessible directly to the users via standard POSIX-like and/or HEP protocols (ie SRM/GridFTP/xroot).
  - Except one site where data is “packaged” and needs user transformation

Survey: Archive reliability

- Observed data losses ranging from O(10^{-8}) to O(10^{-4}) bytes lost / bytes written / year.
  - (or... from 10MB up to 100GB lost/PB/year !!)
  - Big sites: O(10^{-8}) to O(10^{-7})
  - 1 site: 0% loss (redundant data)
  - 1 site: 12% (human error during migration)



Survey: Archive protection / audit

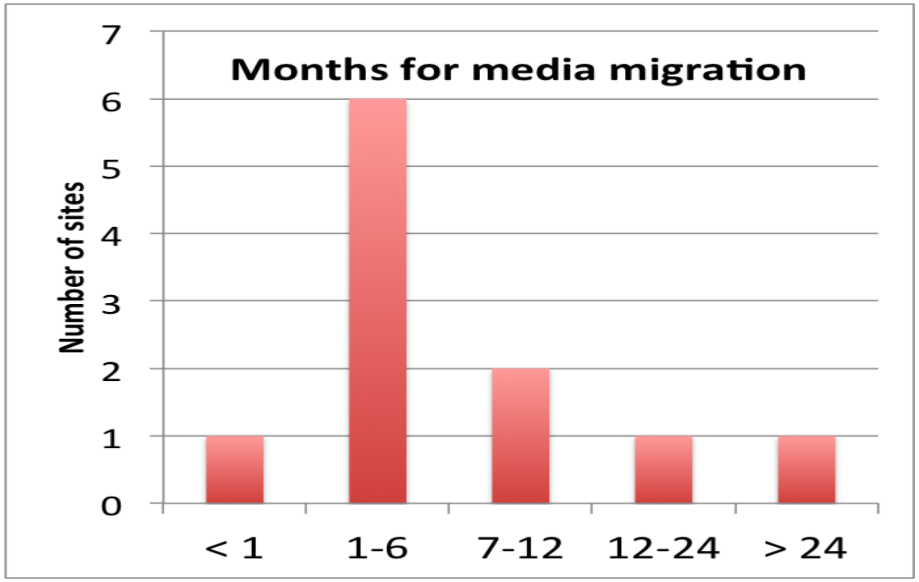
- Most sites (11) use checksums (adler32, md5)
- Most sites (10) use file-level replication for “important” data. No RAIT used
- One (large) site flips the R/O tab once tapes get filled

- (Background) media scans done by half of the sites.
  - All tape contents scanned (2 sites)
  - partial scans (begin/end/random samples) (5 sites)
- Survey shows correlation between higher data loss and no media scanning!

- Data loss notification is manual in all cases
  - Sites contact the owner or the VO/community … via e-mail or issuing a ticket (WLCG)
- Recovery after data loss is always left to the affected user/community
  - No site-to-site recovery procedure even if replicas exist elsewhere.

Survey: Archive migration

- Most sites (13) have media migration experience.
- Migration interval in line with tape hardware cycles: 3-5 years, duration in function of archive size



- Manpower involved for most of sites <1FTE (except larger sites)

- Two sites require users to select what needs to be migrated -> all others: everything!
- 50 % of the sites report integrity problems -> data loss detected during migration.
- 2 sites report that they **won't** be able to bring forward their complete archive to newer storage generations (lack of funding/resources)

Survey: Summary

- Archiving has become a reality by fact rather than by design (MSS==archiving system)
- Often no clear understanding, SLA or agreement of how long archived data should live.
  - Often no funding commitment across complete preservation period.
  - Data owners are currently active experiments, many about to terminate, who takes over?
- Bit rot implying data loss is a reality. Missing QoS or detailed SLA’s defining acceptable data loss rates. Users seem to expect “no data loss”
- Redundancy of data is a way for reducing data loss
  - Intra-site for small / legacy experiments with small footprint
  - Inter-site for larger ones (LHC et al)
- Regular archive audits help improving reliability & reducing migration troubles
  - “Cold” data growing and can’t be left alone
- No inter-site replica recovery; sites rely on users/VO’s for this.
  - Is this optimal?
  - Complexity is high: ~ to (# VO’s) x (Storage systems) x (Sites).

Conclusions and next steps

- Survey gives good overview of HEP bit-level archiving
  - Mostly encouraging results with room for improvement
  - Redo survey outside HEP (large scientific communities) and compare results?
- Collect, document and share best practices at forthcoming HEPiX meetings
  - Archive protection and audit
- Tackle automation of data recovery for distributed archives
- Contributors welcome
  - w3.hepix.org/bit-preservation
  - bit-preservation@hepix.org