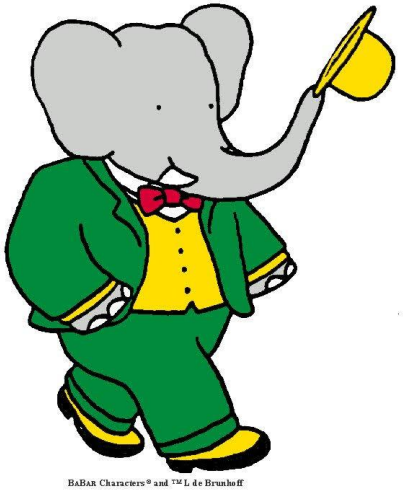
BaBar's Experience with the Preservation of Data and Analysis Capabilities



Marcus Ebert
BaBar Computing Coordinator

on behalf of
BaBar

BaBar status

- BaBar stopped data taking in 2008, anticipated to do data analyses until 2018
  - but still actively doing analyses (local, no Grid usage)
    - 223 active authors from 14 countries
  - 27 new analysis publications since 2018 (more than 60 incl. conference proceedings)
  - 3 new analysis publications in 2023 so far

- beginning of 2021: support for infrastructure at SLAC finally stopped
  - support extended from 2018 to beginning of 2021

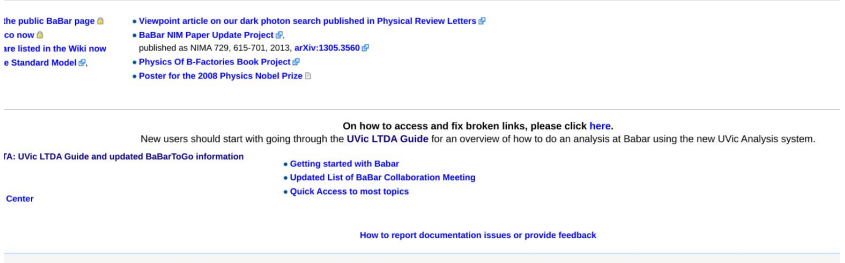
- everything needed to be moved away from SLAC to be able to continue
  - very tightly integration of SLAC services and BaBar services, grown over years

What should be preserved?

No of records displayed are: 28

DOC #	BAIS Title	Submitted	Submitted To
NOTE-1572 …	Corrections to EMC Timing		
NOTE-1571 …	The Care and Feeding of BaBar		
NOTE-1570 …	EMC radFET radiation monitoring system		
NOTE-1569 …	Performance test and comparison of two storage management solutions using BaBar software framework at the Italian Tier-A		
NOTE-1568 …	Performance test and comparison of two storage management solutions using BaBar software framework at the Italian Tier-A		
NOTE-1567 …	Photon Cluster Calibration and Photon Energy Resolution using \Sigma \to \Lambda\gamma Decays		
NOTE-1566 …	Variations in Muon Energy Deposition in the BaBar EMC		
NOTE-1565 …	A search for $B^{+} \rightarrow \tau^{+} \nu_{\tau}$		
NOTE-1564 …	Estimation of \pi^0 detection efficiency using ISR process \gamma\pi^+\pi^-\pi^0		
NOTE-1563 …	Interaction Point Aperture Scan		
NOTE-1562 …	Studying Nonlinear Nonuniformity in the BaBar EMC Crystals		
NOTE-1561 …	New Cluster Splitter/ Photon Finder for the BABAR EMC		
NOTE-1560 …	New Cluster Splitter/ Photon Finder for the BABAR EMC		
NOTE-1559 …	Noise Hit Rate in the AToM Chip		
NOTE-1558 …	Measurement of the q2 dependence of the Hadronic Form Factor in D0 --> K- e+ nu_e decays.		
NOTE-1166 …	Estimation of \pi^0 detection efficiency for low momentum		
NOTE-1165 …	DchOprMonApp and DchOprMiniApp User's Manual		
NOTE-1164 …	The e+ e- --> 3(pi+pi-), 2(pi+pi-pi0) and K+K- 2(pi+pi-) Cross Sections at Center-of-Mass Energies from Production Threshold to 4.5~GeV Measured with Initial-State Radiation		
NOTE-0591 …	Fluorine studies on Babar RPCs		
NOTE-0590 …	HepString Migration		
NOTE-0589 …	EMC Crystal Nonuniformity in the BaBar Monte Carlo		
NOTE-0588 …	Absolute photon energy calibration using pi0's		
NOTE-0587 …	Simulation of non-linear distortions in EMC electronics response		
NOTE-0586 …	LST_FE: a 64 channel front end card with self diagnostic features based on the ALTERA NIOS processor		
NOTE-0585 …	Using the BaBar Publications Database		
NOTE-0584 …	Report of the Study Group on SVT Layer 4 Detector Current Increase		
NOTE-0583 …	A Measurement of Photon Energy Resolution with $D^{*0} \to D^0 \gamma$ Decays		
NOTE-0582 …	Absolute photon energy calibration using pi0's		
NOTE-0581 …	Dependence of the Babar calorimeter resolution on crystal lightyield		
NOTE-0580 …	LST Numbering and Labeling System		
NOTE-0579 …	Performance of the BaBar Electromagnetic Calorimeter with Trickle Injection		
NOTE-0578 …	Test		
NOTE-0577 …	Report of SVT operation during 2002 shutdown		
NOTE-0576 …	The Future of Noneventstore Databases in BaBar		
NOTE-0575 …	A systematic study of radiation damage to large crystals of CsI(Tl) for the BaBar detector		

04/04/2023	Stud
03/23/2023	Searc
02/01/2023	Searc



Physics data

- all data in root files
  - data files and conditions db
  - data access by streaming via xrootd
- metadata in mysql
  - events in which root files are in which collections/skims/run periods/...
  - bad events not to use for analyses
  - file checksums,...
- BaBar used different sites and had already the concept build into its db, e.g. which site has which data and db was duplicated to each site
  - one of the sites is GridKa
    - infrastructure exists, e.g. xrootd, mysql server with db

Physics data preservation

copy all files (via bbcp/xrdcp) to new long term destinations for active usage and backup, verify checksum, and mark as available in the db (backup db too)

- GridKa offered to store data and MC files from the latest processing run (AllEvents, skims, conditions db,...) for active usage via xrootd
- GridKa also continues to host mysql instance with the metadata db
  - xrootd and mysqld at Gridka managed by BaBar
- CC-IN2P3 hosts a second copy of all BaBar data since a long time, incl. raw data, as backup (not for active usage) and agreed to continue to do that
- CERN offered via DPHEP/Open Data Portal to also host a copy of all data

Analysis framework

- BaBar software 32bit, users usually write C++ code and compile their analysis modules
  - does not compile on 64bit-only systems
- depends on older software releases, e.g. perl, xrootd,...
  - latest verified system: SL6.3, gcc 4.4.x, kernel 2.6,...
- all software code under a BaBar specific software-root directory
- BaBar built an own cloud, put into production in 2012 (until end of support at SLAC)
  - VM based on verified OS and software
  - cloud isolated from rest of SLAC, e.g. no central SLAC user management
  - software framework and OS frozen, no changes allowed
  - software framework mounted via NFS

Using VMs and a well defined structure for the software frameworks made it relatively easy to preserve possibility to do analyses.

Analysis framework preservation

archive and copy whole directory tree together with VM images to long term storage and make it usable from there

- easy task (copied all over to UVic which hosts new analysis system)
- but details are not...
  - some links, hardcoded path in source code, scripts,... not using relative paths but absolute SLAC directories
    - mount NFS under the same structure as it was, using /afs/slac....
  - some even point to user directories ($HOME, testing areas,...)
    - long term production tasks run by single users on their own accounts, not on general production accounts
    - users tasked with patching did so in their own area, and compiled/linked from there - dynamically linked so libraries used are in those testing areas still
  - issues only found when things broke after moving the framework out of its initial environment (good exercise to do that from time to time to verify integrity)

Do not run production tasks on personal accounts but role-based accounts!

Analysis framework preservation

In addition:
Since BaBar uses VMs already, create a VM system for users to run on any system.

Two images: OS as used on the central analysis system, second image contains software framework (latest release only, same structure as in the central system).

Users can write, compile, and test their code, and run over any data/MC files when internet connection is available on any machine that supports VMs, including their own laptop.

Accessing the BaBar dataset via the BaBar Associates open-access program

Documentation

- different systems used:
  - html web pages: in AFS within well defined directory structure, r/w rights via ACL, every BaBar user had a SLAC account; edit html files directly in AFS
  - Wiki: added ~2012 to have self contained system editable by anyone in the collaboration via web browser

- html web pages: visible to public or specific groups via .htaccess files, difficult to maintain content

- Wiki: visible only to BaBar members, easy to maintain content

Documentation preservation

copy content to new web servers (for html pages and for wiki)

Easy to do, but... issue again:

- often absolute URLs were used for links instead of relative paths
- some content dynamically created through db queries

- change URLs in html files relatively easy when keeping main structure the same
- db content not accessible and probably will never be - content in SLAC specific Oracle databases
  - created static copies of most pages and content while db access was still possible
    - missed a few...
- having new content in html pages difficult
  - no user accounts on new web server system
  - on new analysis system only for active analysts get accounts

Freeze html content (pages are outdated) and have it no longer available to the public

Documentation preservation

copy content to new web servers (for html pages and for wiki)

Easy to do, but... issue again:

- often absolute URLs were used for links instead of relative paths
- some content dynamically created through db queries

- change URLs in Wiki more complicated
  - content not stored in plain files but in mysql database
  - SLAC IT agreed to have on their web server redirects for BaBar URLs to the new server
  - people should change URLs manually when they come across links that go to SLAC

Wiki became main documentation for BaBar,
old html pages for historic purpose only,
new single public page for general information available.

Collaboration tools

- many different systems needed for the management of a collaboration and to have a communication between members
  - mailing lists
  - meeting pages
  - member lists
  - analysis management and documentations
  - review system for publications and talks
  - communication platform (Hypernews)

- all systems were fully integrated into SLAC central systems and links between the systems
  - people’s database
  - UNIX based authentication and ACLs used to access information
  - systems linked between each other (members database, analysis management system, working groups, mailing lists, Hypernews, email accounts)
  - BaBar specific scripts to query different db to display dynamic information
  - information, incl. binary data like pdf files, in different Oracle databases....

Collaboration tools

- many different systems needed for the management of a collaboration and to have a communication between members
  - mailing lists
  - meeting pages
  - member lists
  - analysis management and documentations
  - review system for publications and ta
  - communication platform (Hyper
- all systems were f egrated into SLAC central systems and links between the system
  - atabase
  - pased authentication and ACLs used to access information
  - systems linked between each other (members database, analysis management system, working groups, mailing lists, Hypernews, email accounts)
  - BaBar specific scripts to query different db to display dynamic information
  - information, incl. binary data like pdf files, in different Oracle databases....

NO TIME TO COVER HERE UNFORTUNATELY

Status summary

Analyses can be done at scale at a new analysis system at UVic or on an own computer based on available VM images.

Data is accessed remotely at GridKa via xrootd streaming and metadata queried remotely at GridKa too.

Documentation available in a self-contained Wiki.

Archival systems for
documentation and communications are at UVic and
for analyses information at INSPIRE.

Active systems for communications and meetings use CERN egroup and Indico, and Caltech based mailing lists.

Active analyses systems use Google drive/sheets/docs.

Summary

- Physics data preservation and open access alone doesn't help for future possibility of doing data analyses
  - one needs to **preserve physics data, analysis framework, and documentation**
  - other tools also important if the collaboration still needs to function at that stage, e.g. doing new analyses

- issues when infrastructure too integrated into local systems
  - centralized and "non-free" databases with data of multiple groups
  - running services based on local accounts (Unix accounts)
  - important services running on personal accounts (cron jobs)
  - how information and code is written (absolute paths...)
- using systems that can be ported to other places (opensource) helps a lot

Conclusion

Data and analyses preservation can be done,
but planning early for it can help a lot,
especially when choosing systems/formats/conventions while an experiment runs.

BaBar went through this and setup a new computing infrastructure independent of original host laboratory - and it works.

Systems and organisations available to help with long term archival and preservation
DPHEP, OpenData portal, Inspire,...

Big Thanks
to the
GridKa, CERN, CC-IN2P3, Inspire, Caltech, and UVic HEPRC groups!

Collaboration tools

- SLAC based mailing lists ---> Caltech mailing lists
  - only created what is still needed

- old meeting agendas were HTML pages, registration based on SLAC systems

---> switch to use CERN Indico

- Hypernews was deeply integrated into SLAC
  - sending emails for posts to SLAC emails, notify SLAC systems in case of issues, people joining need SLAC UNIX account,... - but all content of posts in text files

---> moved Hypernews out of SLAC, made read-only, and removed any mailing feature -> still readable and archive of any communication happened in the past

---> replacement: CERN egoups
  - also nicely integrated with CERN Indico for accessing BaBar meetings

Collaboration tools

- Analysis documents, notes, and Analysis metadata
  - old content archived to INSPIRE
  - new documents will be added for long term preservation too

<u>new system for active analyses and management:</u>

- Google drive folder for each analysis
  - for documents documents and other informations
- Google sheets for metadata of each analysis
- review done using CERN egroups (each analysis has its own)
- specific folders for SpeakersBureau, PublicationBoard,...