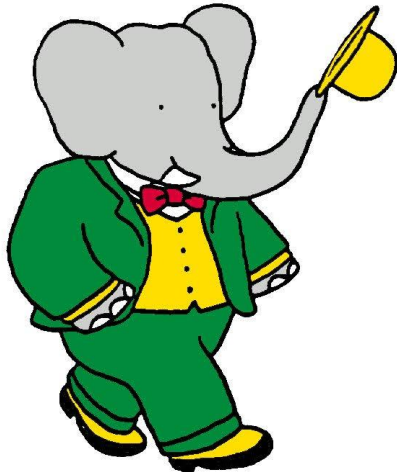


Preservation of Data and Analysis Capability at BaBar



BaBar Characters® and ™ L. de Bruijn

Marcus Ebert
BaBar Computing Coordinator

BaBar status

- BaBar stopped data taking in 2008
 - but still actively doing analyses (local, no Grid usage)
 - 4 publications in 2021, 5 in 2022 so far, just paper #600 was accepted for publication
- since 2019: very minimal support for BaBar infrastructure at SLAC
 - reduced to some support that involves entering machine room
- beginning of 2021: support for infrastructure at SLAC stopped

BaBar just went through all the preservation steps

- everything needed to be moved away from SLAC and can't depend on SLAC infrastructure anymore
 - very tightly integration of SLAC services and BaBar services, grown over years

What should be preserved?

- Data
 - collected Data, MC events (signal and generic)
 - Metadata

What should be preserved?

- Data
 - collected Data, MC events (signal and generic)
 - Metadata
- Analysis framework
 - analysis software, OS, specific version of tools
 - perl, python, xrootd, root libs,...

What should be preserved?

- Data
 - collected Data, MC events (signal and generic)
 - Metadata
- Analysis framework
 - analysis software, OS, specific version of tools
 - perl, python, xrootd, root libs,...
- Documentation to make use of the analysis framework
 - software description
 - data available and how to access (skims, produced MC samples,...)

What should be preserved?

- Data
 - collected Data, MC events (signal and generic)
 - Metadata
- Analysis framework
 - analysis software, OS, specific version of tools
 - perl, python, xrootd, root libs,...
- Documentation to make use of the analysis framework
 - software description
 - data available and how to access (skims, produced MC samples,...)
- Collaboration tools
 - discussion places, email list, meeting software, member database,...

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 the db, e.g. which site has which data and db was duplicated to each site
 - one of the sites is GridKa
 - infrastructure, e.g. xrootd, mysql server with db

Data preservation

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)

Data preservation

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
- GridKa also continues to host mysqld with the metadata db
 - xrootd and mysqld at Gridka managed by BaBar
- IN2P3 hosts since a long time a second copy of all BaBar data, 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
- depends on older software releases, e.g. perl, xrootd,...
 - latest verified system: SL6.3, gcc 4.4.x, kernel 2.6,...
- BaBar built an own cloud, put into production in 2012
 - 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

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

Analysis framework preservation

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,...)
 - production tasks run by single users long term 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 in those testing areas still
 - issues only found when things broke after moving the framework out of its initial environment

Do not run production tasks on personal accounts!

Analysis framework preservation

In addition:

Since BaBar uses VMs already, create a VM 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.

possibility to do analyses: BaBar members and associates (open to anyone)

'BaBar Associates' open-access: anyone can join, full access to communications and documentation tools, analysis for publication to be done within BaBar framework, e.g. going through the full review process ---> *to be reviewed at one point*

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

documentation preservation: copy content to new web servers
(for html page and for wiki)

Documentation preservation

documentation preservation: copy content to new web servers
(for html page 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
 - SLAC infrastructure grew out of BaBar IT infrastructure
 - created static copies of most pages and content while db access was still possible
- 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

documentation preservation: copy content to new web servers
(for html page 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 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

- 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 too for long term preservation

new system for active analyses and management:

- [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,...

Status summary

Status summary

Analyses can be done at scale at a new analysis system at UVic or an own system based on available VMs can be used.

Status summary

Analyses can be done at scale at a new analysis system at UVic or an own system based on available VMs can be used.

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

Status summary

Analyses can be done at scale at a new analysis system at UVic or an own system based on available VMs can be used.

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

Documentation available in a self-contained Wiki.

Status summary

Analyses can be done at scale at a new analysis system at UVic or an own system based on available VMs can be used.

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 at UVic and
for analyses information at INSPIRE.

Status summary

Analyses can be done at scale at a new analysis system at UVic or an own system based on available VMs can be used.

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 at UVic and
for analyses information at INSPIRE.

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

Status summary

Analyses can be done at scale at a new analysis system at UVic or an own system based on available VMs can be used.

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 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 tools.

Summary

- Data preservation and open access alone doesn't help for future possibility of using the data
 - one needs to **preserve data, analysis framework, and documentation**
 - other tools also important if the collaboration still needs to function at that stage, e.g. doing new analyses
 - for that reason, anyone can use the data, but in a “controlled way” when aiming for publication: [‘BaBar Associates’ open-access program](#)
- large issues for 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.

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 just went through this and setup a new infrastructure away from SLAC - and it works.

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 just went through this and setup a new infrastructure away from SLAC - and it works.

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

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 just went through this and setup a new infrastructure away from SLAC - and it works.

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

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