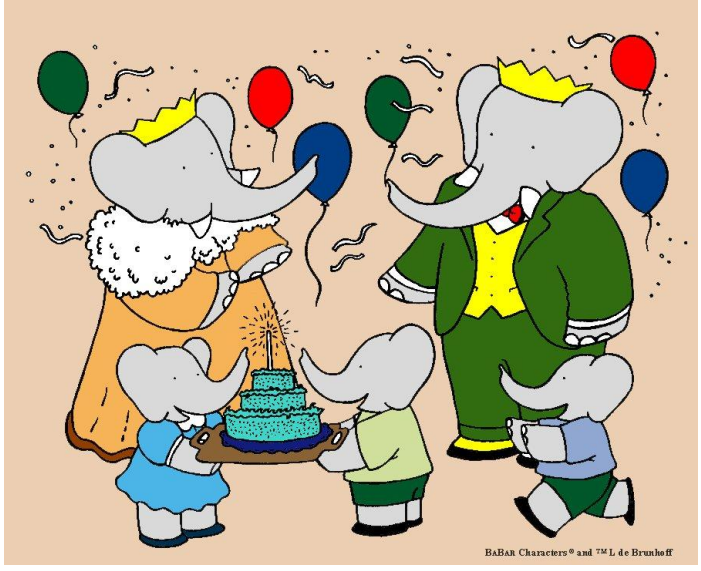
Data Preservation and Open Data at the B-factories

Marcus Ebert
University of Victoria

- Why data preservation?
- What should be preserved?
- How to do preservation.
- lessons learned



Why Data Preservation?

- new ideas how to do an analysis
  - (better software, particle ID, reduced systematics,...)
- new ideas what to look for
- new theories to check
- cross-check results of other experiments
- unique data sets may exist
- teaching
- ...

B-factories

BaBar

- formed 1993
- data taking at SLAC from 1999-2008
  - stopped 16 years ago...
- beginning of 2021 support of data access and computing at SLAC ended
- new computing infrastructure outside of SLAC working since beginning of 2022

Belle

- formed 1993
- data taking at KEK from 1999-2010
  - stopped 14 years ago...
- followed by Belle-II at KEK
- computing infrastructure still at KEK

What should be preserved?

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

Data preservation

copy all files to new long term destinations for active usage and backup

Data preservation

copy all files to new long term destinations for active usage and backup

BaBar

- GridKa offered to store data and MC files from the latest processing run (AllEvents, skims, conditions db,...) for active usage
  - also continues to host metadata db
- IN2P3 hosts since a long time a second copy of all BaBar data, incl. raw data, as backup and agreed to continue to do that
- CERN offered via DPHEP/Open Data Portal to also host a copy of all data

Belle

- Belle data continuous to be stored at the KEK Computing Centre
  - main computing centre for Belle-II
  - at least until Summer 2024

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,...

Analysis framework

- BaBar software 32bit, users usually write C++ code and compile their analysis modules
  - does not compile on 64bit only infrastructure
  - depends on older software releases, e.g. perl, xrootd,...
    - latest verified system: SL6.3, gcc 4.4.x, kernel 2.6,...
      - frozen software environment
      - OS unsecure

-----> **Virtualize framework and run everything from within Virtual Machines**

- Belle software frozen but ported to CentOS7
  - Summer 2024 also EOL for CentOS7
  - **tools developed to use Belle-II software for data analyses**

BaBar analysis framework preservation

- everything is in a single directory tree and in VM images
- preservation is an easy task
  - (copied whole directory tree to UVic which hosts new analysis system and setup VMs based on the cloudscheduler 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 - binaries copied over but dynamically linked libraries stayed in those testing areas
  - issues only found when things broke after moving the framework out of its initial environment

BaBar analysis framework preservation

- everything is in a single directory tree and in VM images
- preservation is an easy task
  - (copied whole directory tree to UVic which hosts new analysis system and setup VMs based on the cloudscheduler 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 - binaries copied over but dynamically linked libraries stayed in those testing areas
  - issues only found when things broke after moving the framework out of its initial environment

Do not run production tasks on personal accounts!

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,...)

BaBar 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

BaBar documentation preservation

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

BaBar 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 (easily) accessible - content in SLAC specific Oracle databases
  - 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 UVic web server system
  - on new UVic analysis system only active analysts get accounts

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

BaBar 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.

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,...

BaBar 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....

BaBar 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

BaBar collaboration tools

- Analysis documents, notes, and Analysis metadata
  - old content archived at 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,...

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,...

Open Data

- making data openly available possible but not useful by itself
- to make use of the data one also needs
  - Analysis framework
  - Documentation
  - Communication with collaboration members

Access to BaBar framework: new analysis system at UVic, BaBar-To-Go at home

‘BaBar Associates’ open-access:

- anyone can join (== data access for anyone)
  - full access to communications and documentation tools and archives
  - analyses for publication to be done within BaBar publication framework
    - e.g. going through the full review process
  - https://babar.heprc.uvic.ca/www/join_BaBar.html

Lessons learned

- don’t run production tasks on personal accounts
- relative addresses/path whenever possible
- make use of open source tools
- data preservation includes at least data files and metadata, analysis framework, and documentation
- early planning of long term preservation behind an experiments data taking period helps a lot
- Organizations out there to help with preservation
  - CERN: long term storage of data via Open Data project: https://opendata.cern.ch/docs/about
  - DPHEP: https://dphep.web.cern.ch/about
    - **DPHEP Global Report 2022 h**ttps://link.springer.com/article/10.1140/epjc/s10052-023-11885-1
  - INSPIRE: https://inspirehep.net/

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.

Systems and organisations are 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.

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

Thank You!