The Development and Use of Public Data Bases: It's Complicated!

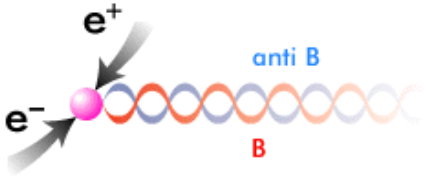
BaBar and the HEP Panorama on Matters of Data Preservation and Public Access

Concetta Cartaro
SLAC

Progress on Statistical Issues in Searches
SLAC, June 4th, 2012

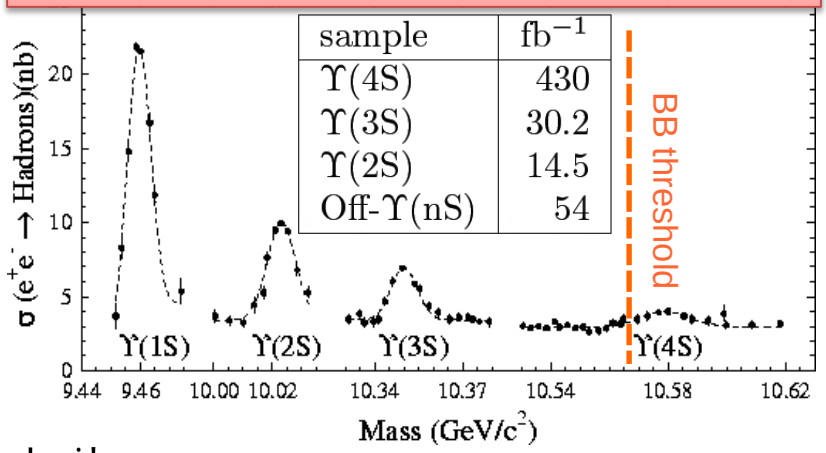
INTRODUCTION TO BABAR

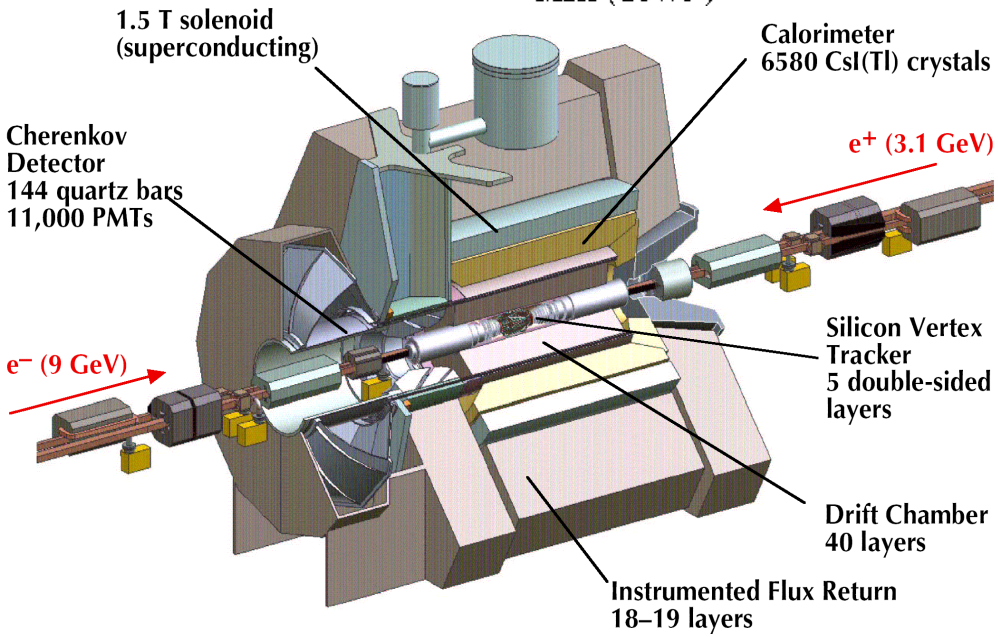
- The PEP-II asymmetric e^+e^- storage ring at SLAC is a flavor factory more than a *B*-factory:
  - 470 x 10^6 *B* anti-*B* pairs
  - 690 x 10^6 *c* anti-*c* pairs
  - 500 x 10^6 $\tau^+\tau^-$ pairs



Center-of-mass (CM) energy corresponds to the mass of the Y(4S) resonance, at *B anti-B* production threshold. These events are used for the study of B-meson decays, CP violation, and B^0 *anti-*B^0 mixing. Data samples collected at the Y(3S) and Y(2S) resonances in 2008 are used for bottomonium studies and for dedicated new-physics searches. For each Y(nS) resonance (n = 2, 3, 4), "off-resonance" sample was collected for studying continuum e+e– → q anti-q events, where q is a u, d, s, or c quark. Both on and off-resonance samples are used for charm, τ, two-photon, and hadronic physics. The physics program includes also light Higgs and dark matter searches.

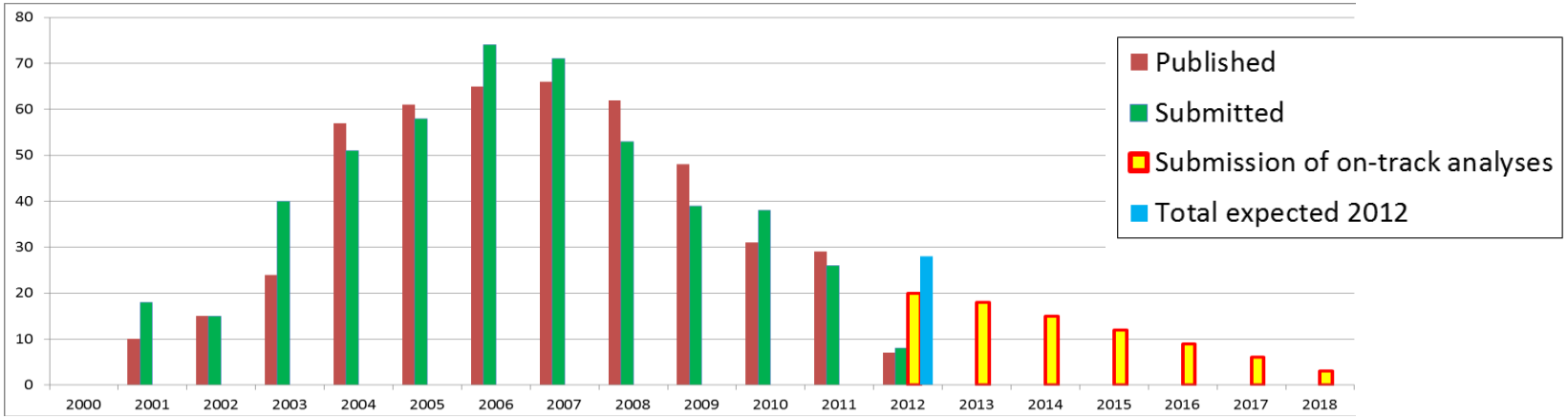
Y(4S) = b anti-b quark bound state @ 10.58GeV



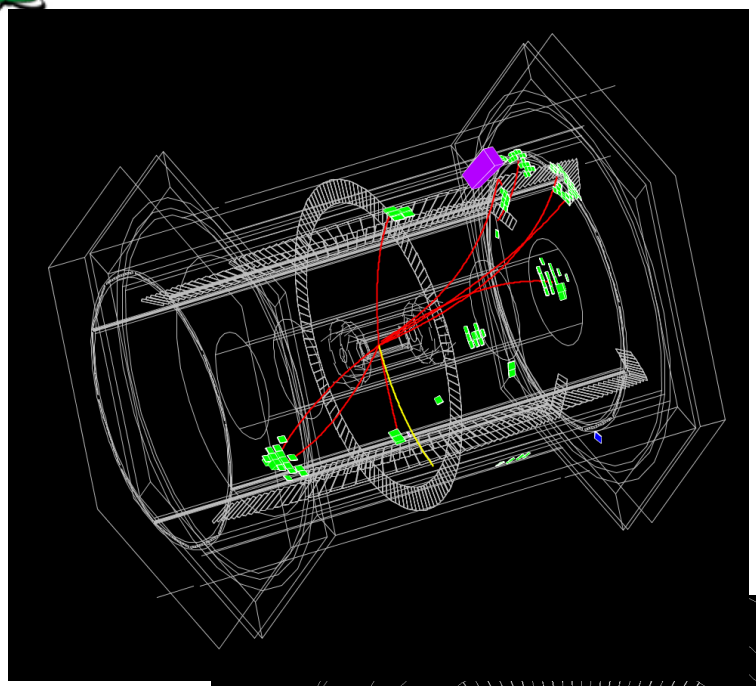


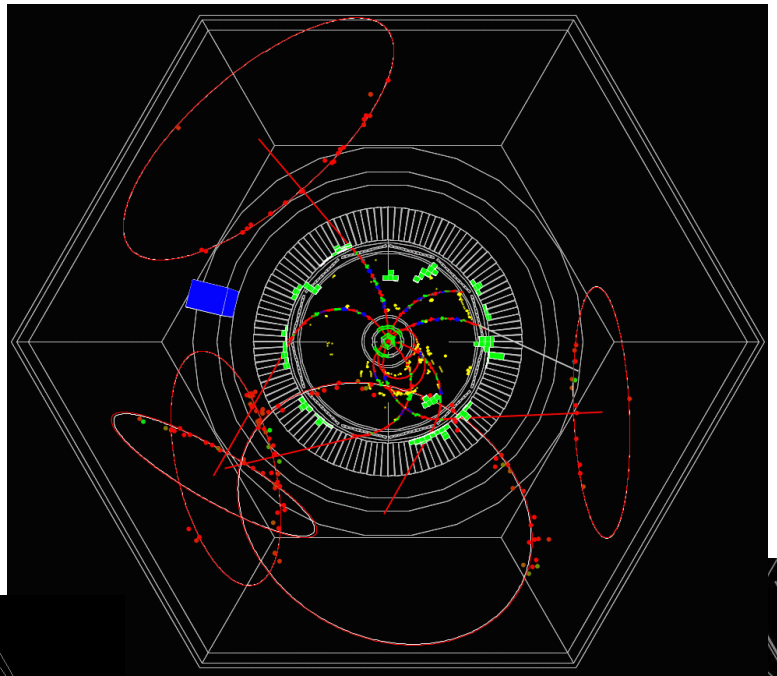
BABAR FACTS

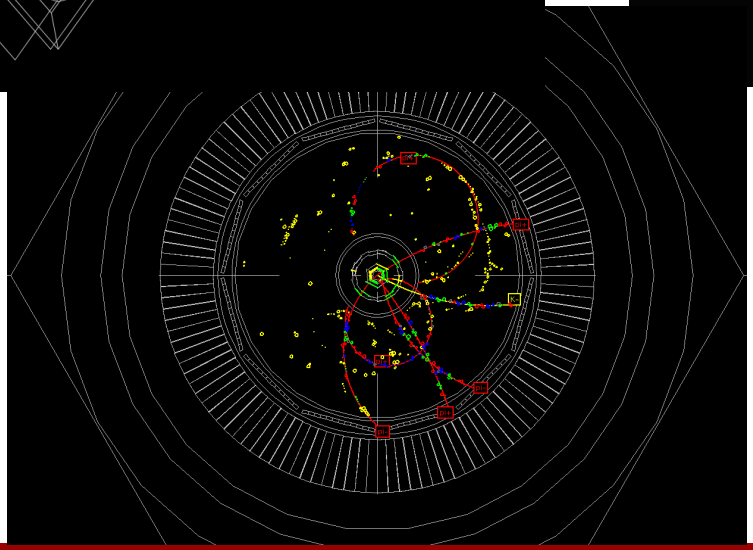
- BaBar has collected data from Oct 22nd 1999 to Apr 7th 2008
  - 800TB of raw data, 1.2 PB from the last data reprocessing
  - 476 published papers to date
  - 84 on track analyses
    - Plus ~60 with lower publication probability (generally lacking manpower)
    - Possibilities for new previously unforeseen analyses including discovery analyses
- 364 members (not including associates) from 73 institutions in 12 countries
  - 23 people joined BaBar as Members since 1 Jan 2010; 11 since 1 Jan 2011
  - 31 Associates joined since 1 Jan 2011; 7 new Associates since 1 Jan 2012
    - 3 new Undergrad started the association procedure last week

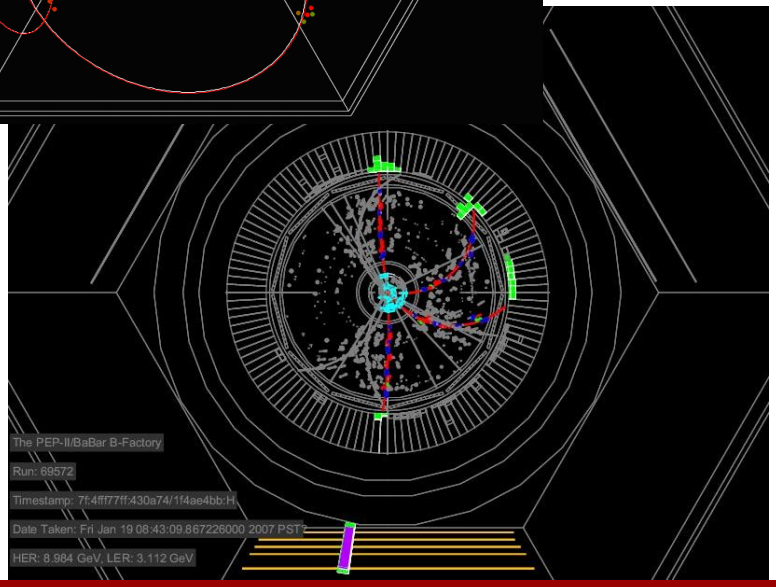


BABAR EVENTS









BABAR EVENT STORE

http://root.cern.ch/

- The raw data from collision events are collected in *runs*, that are then processed in order to calibrate and reconstruct the single events.
  → Reconstructed events are stored in collections in ROOT format, independent from the original event format or production system.
- The collections are stored on tape and on disk at SLAC and, partially, at several other locations, Tiers, supporting BaBar analysis activity in US, Canada and EU.
- Bookkeeping databases (Oracle and MySql) are used to keep track of every single collection, from file info to quality information.
  - Trivia: BaBar bookkeeping keeps track of 1.35×10^6 collections in 2×10^6 files and 2PB of size (only latest reprocessing).
- The software is almost all C++ Object Oriented, with Tcl scripts to drive the C++ framework. Python and Perl are used as well in some tools.
  - Trivia: BaBar code contains more than 2×10^6 lines of code (not including users code) and our releases are rebuilt every night on all the supported Linux and Sun OS's.

What I Was Asked

- Why BaBar data is not public ?
  - Beware, there is not one simple answer!
  - Like all the other HEP experiments BaBar data is managed by the collaboration and, like most, is kept private. But then we started to talk about long term data preservation.
    - It took a solid decade to BaBar to become a mature experiment, to fully grasp the complexity of the data, of the detector, of the code, and of the physics and take on all the challenges making it the successful experiment it is. How can a non-*BaBarian* hope to dig out good physics from this?
  - But still, do not confuse preservation and public access. They are different …Or not ?

    → Can open access drive the data preservation?

On the Other Hand

- Nobody asked (is that an excuse?)
  - The DOE, unlike NASA, has no requirements on public access. It only requires a statement on the data preservation plan. And “no plan” is a plan.
  - The other international funding agencies never had a policy (stringent) on public access. After all the HEP projects are very expensive and have to be exploited to their maximum for the benefit of the Collaboration itself and the funding agencies.

    → The cost (public money) of such experiments is exactly why they deserve to be preserved, open access or not

A Convoluted Problem

- HEP experiments work in a highly competitive mode for the majority of their lifetime driven by the huge wealth of physics contained in the data.
- And when the discovery potential of the data is exhausted?
  - Does it even happen? What about new physics ?
  - Move on to the next generation experiment.
- A recent problem:
  - Next generation may happen in 20 years if ever …
    - Will CDF data be ever superseded ? Probably not.
  - Start thinking about data preservation
- Next problem: cost.

A Matter of Resources?

- Preservation means cost.
- And public data? Even more costly (maybe… ?)
  - Very few believe that good results can be obtained by non collaborators without the Collaboration support (from the documentation – lacking of course – to the review process)
- And even if we were able to deliver detailed documentation, how can we trust the results?
  - Is a disclaimer stating that the results obtained with BaBar (or any other HEP project) public data are not endorsed by the Collaboration enough?
  - Many feel very strongly about the data and their misuse. To the extent we don't even trust ourselves: Blind analysis

JUST SCRATCHING THE SURFACE

- Intellectual property on data and/or code.
  - Clearly stated in some (not all) Memorandum of Agreement between BaBar institutions and SLAC.
    - Changes of policy are announced Collaboration wide. No answer means agreement.
    - Use case: Institutions that asked permission to use some of our data for education. It was, of course, granted.
- Internal docs and code.
  - Internal documents are sometimes written with little regard to the form, they are practical, technical documents full of jargon terms with the sole aim of describing the analysis for internal review.
    - If public, some could consider these documents detrimental to their career.
    - Also in this case, provided the authors agreement, the documents could be given to individuals that asked for a specific reason and in the interest of the Collaboration.
- Third party code, licensing and copyright, …
  - In the long term planning we're trying to get away from these has much as possible.
- Computing resources (CPU, disk, …)

PRIVATE, BUT:

- In the case of BaBar, while it is true that the data, the code and the documents are not open, it is also true that anyone can ask to become a "*BaBarian*".
  - If you have an idea, if you want to verify a result, you can ask to become a BaBar associate and have full access to all our resources, data, code, documents, computing plus the strength of a Collaboration ready to help at any moment, providing support, review, discussion, …
  - Based on an honor system all BaBar members and associates are asked not to share and distribute our internal material unless approved by our review process.

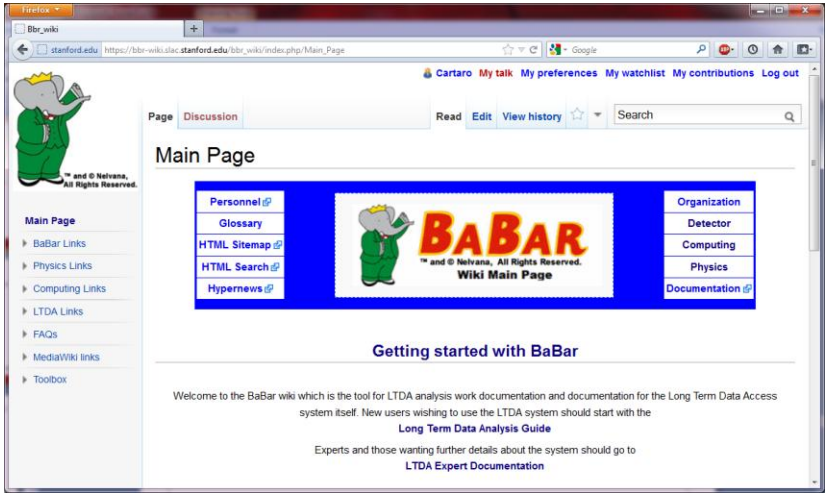
BaBar associates are authors of their own papers only. They can become members (and full authors) after one year or after their first publication whichever comes first. Being a member means also taking an active role in the Collaboration fulfilling duties like service tasks and participating in the review process of the results.

BABAR BEYOND 2012

- Insure the ability to do analysis on the BaBar data beyond 2012 and until at least 2018 preserving:
  - Data, conditions and calibrations, releases and tools, databases, capability of running production and user jobs
    - This means that in 5 years from now it will be possible, for example, to add a new decay mode, produce the MC events and the relevant skims, and perform a completely new analysis developing new selection code, fitting procedures, etc.
  - Documentation
- Close BaBar Framework into a frozen environment
  - Last validated OS enclosed in a virtualization layer running the BaBar code
  - Freeze the environment, not the Framework!
- BaBar on the clouds
  - Virtualization can provide virtual hardware support on which we can run virtual images with the wanted OS
    - However a VM connected to a network is not different from a physical machine and running old OS poses a security threat

Documentation Working Group

- Strong push toward documentation clean up, ease of access, and clarity.
- All most used and fundamental info are being checked, updated and moved to a Media Wiki server, the *BABAR WIKI*
  - Detector pages and other pages that will supposedly never change again will be left in their original location
- Created a Documentation Working Group to coordinate the migration effort aided by an advisory committee
  - There are 10 official members in the DWG but we promote the migration to the wiki as a Collaboration effort
  - Many new students joined the effort but the input from senior members of the Collaboration is fundamental
- The effort needed is not trivial



New BaBar Public Page

http://www-public.slac.stanford.edu/babar/

BABAR

SLAC NATIONAL ACCELERATOR LABORATORY

- Home
- Purpose of BABAR
- BABAR Physics
- How BABAR Works
- BABAR Publications
- All BABAR News
- Images, Video, & More
- Organization
- SLAC Home
- BABAR Internal Page
- Site Index

The BaBar Experiment

Welcome to the BABAR public web site. BABAR is a particle physics experiment designed to study some of the most fundamental questions about the universe by exploring its basic constituents - elementary particles. The BABAR Collaboration's research topics include the nature of antimatter, the properties and interactions of the particles known as quarks and leptons, and searches for new physics. We invite you to explore the site and learn about the BABAR detector, our research, and the physicists who perform it.

Recent News

- **Particle-Physics History on Display**
  May 9, 2012
  BABAR's innermost system, the Silicon Vertex Tracker, gets a new life as a unique exhibit of particle-detector design and construction.
- **Hunting Dark Photons and Higgs with BABAR: Science Highlight and Symmetry Magazine Article**
  May 2012
  Light dark photons? Dark Higgs bosons? Scientists look for signs of these weird-sounding particles in data from BaBar—an experiment designed to explain a completely different mystery.
- **BABAR Members Greeted Warmly at Wintery Conference**
  March 1, 2012
  Four members of the BaBar collaboration braved frigid temperatures to present the

BABAR Highlights

- **BABAR's role in the 2008 Nobel Prize in Physics, December 8, 2008.** Matter-Antimatter Asymmetry Measurements Put Final Seal of Approval on the Theory of Quarks.
- **BABAR Discovers the Bottom-Most Bottomonium, July 9, 2008.** The "ground state" of $b\bar{b}$ quark pairs is finally detected.
- **New form of matter-antimatter transformation observed for the first time, March 13, 2007.** BABAR finds evidence for mixing in the charm system.
- **BABAR discovers a new massive particle, July 1, 2005.** The Y(4260) is not expected theoretically and its nature is a mystery.
- **BABAR sees direct charge-parity violation, August 2, 2004.** B mesons and their antiparticle partners undergo a radioactive decay at different rates.
- **BABAR announces new result on charge-parity violation, July 23, 2002.** Precise measurement of the parameter sin2β.
- **BABAR establishes charge-parity**

A Wider Landscape

- Size, complexity, cost and timescale of experiments has greatly increased, period between an experiment and its successor has become long and uncertain. Data is too rich and analyses take too long to be fully exploited during the lifetime of the collaborations.
- We are not the only ones who believe their data are precious
  - Collider experiments (ee, ep, pp), computing centers, funding agencies
- A matter of scale:
  - BaBar raw data output rate 530 KB/s (2PB)
  - ATLAS raw data output rate 300 MB/s (15 PB/year!)
  - LSST calibrated data output rate 60000 MB/s every 40 seconds…

ICFA Study Group on Data Preservation and Long Term Analysis in High Energy Physics

A.K.A. DPHEP

- **International Steering Committee**
  DESY-IT: Volker Gülzow (DESY)
  H1: Cristinel Diaconu (CPPM/DESY)
  ZEUS: Aharon Levy (Univ. Tel Aviv)
  HERMES: Gunar Schnell (DESY)
  FNAL-IT: Victoria White (FNAL)
  D0: Dmitri Denisov (FNAL), Gregorio Bernardi (LPNHE)
  CDF: Giovanni Punzi (FNAL), Robert Roser (FNAL)
  IHEP-IT: Gang Chen (IHEP)
  BES III: Yifang Wang (IHEP)
  KEK-IT: Takashi Sasaki (KEK)
  Belle: Hisaki Hayashii (Univ. Hara), Leo Piilonen (Virginia Tech), Yoshihide Sakai (KEK)
  SLAC-IT: Richard Mount (SLAC)
  BaBar: Michael Roney (SLAC/Victoria)
  SLAC: Amber Boehnlein (SLAC)
  CERN-IT: Frederic Hemmer (CERN)
  ATLAS: Fabiola Gianotti (CERN)
  CMS: Guido Tonelli (CERN)
  LHCb: Pierluigi Campana (CERN)
  ALICE: Paolo Giubellino (INFN/Torino)
  CERN/PARSE: Salvatore Mele (CERN)
  CLEO: David Asner (Carleton)
  JLAB: Graham Heyes (JLAB)
  BNL: Michael Ernst (BNL/IT)
  STFC: John Gordon (RAL)

- **International Advisory Committee**
  Jonathan Dorfan (SLAC) (co-chair)
  Siegfried Bethke (MPI Munich) (co-chair)
  Young-Kee Kim (FNAL)
  Hiroaki Aihara (U.Tokio)
  Dominique Boutigny (IN2P3)
  Michael Peskin (SLAC)
  Gigi Rolandi (CERN)
  Alex Szalay (JHU)

Contact Persons for other communities and projects

Fabio Pasian (Trieste) International Observatory for Astrophysics
Sayeed Choudhury (John Hopkins Univ. USA)
Data Conservancy/Blue Ribbon
Adil Hassan (QMU London) DRESNET
Robert Hanisch (STSCI USA) IVOA

ICFA = International Committee for Future Accelerators

http://www.dphep.org/e26/
Warning: affiliations maybe out of date

HTTP://WWW.DPHEP.ORG/

Firefox | ICFA Study Group on Data Preservation ... | www.dphep.org

Study Group for Data Preservation and Long Term Analysis in High Energy Physics

Home | People | Committees | Subgroups | Workshops | Documents | Work Space | Press

DPHEP

ICFA Study Group on Data Preservation and Long Term Analysis in High Energy Physics

New publication May 2012, available here:
Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics

DPHEP@CHEP2012: DPHEP Session at CHEP 2012, New York, May 24 2012

High Energy Physics experiments initiate with this **Study Group** a common reflection on **data persistency and long term analysis** in order to get a common vision on these issues and create a multi-experiment dynamics for further reference.

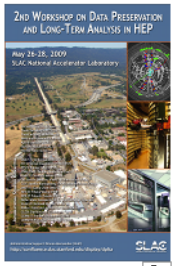
The objectives of the Study Group are:

- Review and document the physics objectives of the data persistency in HEP.
- Exchange information concerning the analysis model: abstraction, software, documentation etc. and identify coherence points.
- Address the harware and software persistency status.
- Review possible fundings programs and other related international initiatives.
- Converge to a common set of specifications in a document that will constitute the basis for future collaborations.

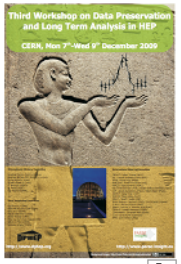
Since August 2009, the Study Group is endorsed by ICFA (International Committee for Future Accelerators).

A series of workshops have been held by the Study Group, access to which can be found using the links below.
The 3rd workshop was preceded by a public **symposium**.
The first DPHEP publication, containing the initial recommendations of the Study Group was released in December 2009.
This was followed by a more comprehesive, **second publication**, in May 2012.

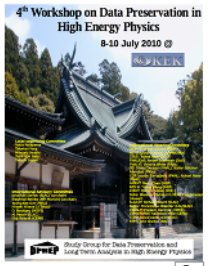
DESY January 2009



SLAC May 2009



CERN December 2009



KEK July 2010

Fermilab May 2011

Latest DPHEP Status Report released on arxiv:1205.4667

CERN Courier, April 2009

DPHEP Mission

- The DPHEP promotes experiment level projects of data preservation in all the major labs in order to avoid loss of data after the collaborations are dissolved or are too small to support their infrastructure.
  - Collect details about the single projects and their needs
  - Define working directions and organization
- The DPHEP also promotes inter-collaboration communication:
  - Sharing experiences
    - BaBar, H1, Zeus, Belle, Bes III, Jade, LHC, …
  - Defining common projects
    - Data preservation models and tools
    - Virtualization and archival infrastructure, data migration and validation, etc.
  - Documentation and Organization
    - Extension and enhancement of documentation by storing figures,
    - n-tuples/root-tuples, and other legacy documents
  - Outreach
    - Standard format tools for sharing data and education

PRESERVATION MODELS

BaBar choice

Preservation Model	Use Case
Level 4: Preserve the reconstruction and simulation software and basic level data	Full potential of experimental data
Level 3: Preserve the analysis level software and data format	Full scientific analysis based on existing reconstruction
Level 2: Preserve the data in simplified format	Outreach, simple training analyses
Level 1: Provide additional documentation	Publication related information search

benefits and cost

DPHEP Meeting Highlights (I)

Hosted by CHEP, New York on May 24th 2012

- CDF data preservation project started.
- LHC experiments now organize regular internal meetings on data preservation.
  - Policies are being discussed and common lines defined
  - CMS has approved a policy on data preservation and open access in March and will release data according to Level 3 three years after data taking.
- H1, Hermes and Zeus at DESY have all started their preservation projects with a strong interaction with INSPIRE. A software validation platform has been developed in order keep the software updated and working on newer platforms.

http://www.chep2012.org/

https://indico.cern.ch/conferenceDisplay.py?confId=171962

DPHEP Meeting Highlights (II)

Hosted by CHEP, New York on May 24th 2012

- All experiments will implement Level 1 through INSPIRE and HEPData. Level 2 is aimed to outreach and will be implemented by most depending on resources.
- Level 3 for public access can cause lengthy discussions within collaborations and requires resources often not available and not initially foreseen by funding agencies.
  - Strict rules on open access and results produced outside the collaborations: everything is “as is”; collaborations don’t acknowledge the results and don’t provide review; collaborators are forbidden to sign papers not originating from within the collaboration (LHCb proposed policy – not yet agreed or endorsed).
- Level3 and Level4 for collaborations are carefully considered and will be carried on at the best of collaborations capabilities.

"With Great Power Comes Great Responsibility" (Uncle Ben)

NATURAL ENVIRONMENT RESEARCH COUNCIL

http://www.nerc.ac.uk/research/sites/data/policy.asp

Freely available without any restrictions on use

NATIONAL INSTITUTES OF HEALTH

http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

wellcome trust

Data sharing

http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm

Maximize the availability of research data

Data management and sharing plan

NSF

http://www.nsf.gov/bfa/dias/policy/dmp.jsp

Science & Technology Facilities Council

http://www.stfc.ac.uk/resources/pdf/stfc_scientific_data_policy.pdf

Data resulting from publicly funded research should made publicly available after a limited period, unless there are specific reasons (e.g. legislation, ethical, privacy, security) why this should not happen.

publicly archived all of its data received from spacecraft project

NASA

http://www.nasa.gov/open/data.html

Social Accountability ?

DPHEP

Sünje Dallmeier-Tiessen, Salvatore Mele

CERN – Scientific Information Service – Open Access Section

DPHEP@CHEP NYC May 24th 2012

FITS Documentation Page

fits.gsfc.nasa.gov/fits_documentation.html

FITS The Astronomical Image and Table Format
Flexible Image Transport System

Home | News | Docs | WCS | Samples | Libraries | Viewers | Utilities | Keywords | Conventions | Resources

FITS Documentation

Descriptions of the FITS format

- FITS Overview - brief history of the origins and evolution of FITS
- FITS Format Primer - introduction to the basic structure of a FITS file
- The FITS Standard - the definitive reference document that defines the FITS format. Version 3.0 was approved in July 2008.
- World Coordinate System (WCS) - documents and software dealing with world coordinate system conventions. The first 3 WCS papers have been approved by the IAU-FW

tDAR

www.tdar.org

tDAR the Digital Archaeological Record

A SERVICE OF DIGITAL ANTIQUITY

Home | About | Features | Why Use tDAR | News | How to Use tDAR | Our Team | Contact Us

Home

WELCOME to the Digital Archaeological Record (tDAR). tDAR is an international digital archive and repository that houses data about archaeological investigations, research, resources, and scholarship. tDAR provides researchers new avenues to discover and integrate information relevant to topics they are studying. Users can search tDAR for digital documents, data sets, images, GIS files, and other data resources from archaeological projects spanning the globe. For data sets, users also can use data integration tools in tDAR to simplify and illuminate comparative research.

Excavations atop Mound A, Shiloh National Military Park, Tennessee.

ACCESS is a core feature of tDAR. tDAR supports broadening the access to a wide variety of archaeological data. Browsing or searching the tDAR repository enables users to identify digital documents, data sets, images, and other kinds of archaeological data for research, learning, and teaching. tDAR enables users to download data files while maintaining the confidentiality of legally protected information and the privacy of digital resources on which a researcher is still working.

PRESERVATION is the other key part of tDAR's mission. tDAR and its parent organization, Digital Antiquity, are dedicated to ensuring the long-term preservation of digital archaeological data. These data document the archaeological record, the efforts of the archaeological and scientific community, and the material and social characteristics of the cultures studied.

GROWTH and IMPROVEMENT are part of Digital Antiquity's strategy for tDAR. Regular enhancements and improvements that incorporate advances in research methods, digital preservation, and technology are

Use tDAR

Search tDAR search

Login | Register | Browse

Explore

What is under Phoenix? Read about the archaeology of 1000-year-old towns and canals under the modern city.

Research

Connecting to our past, tDAR is advancing archaeologists' ability to engage in synthetic and comparative research. learn more ...

Access & Preservation

A network of digital information that is available for you and for generations to come! tDAR provides a means to maintaining the long-term utility and accessibility of irreplaceable primary data in the face of inadequate metadata and rapidly changing technology. learn more ...

Public Access & Education

tDAR is committed to ensuring public access to and use of materials, whether for Section 106

UK Data Archive - HOME

www.data-archive.ac.uk

This website works best if you accept cookies I ACCEPT Tell me more

HELP CONTACT US SIGN UP

UK·DATA ARCHIVE

THE UK'S LARGEST COLLECTION OF DIGITAL RESEARCH DATA IN THE SOCIAL SCIENCES AND HUMANITIES

HOME | ABOUT US | CREATE & MANAGE DATA | DEPOSIT DATA | HOW WE CURATE DATA | FIND DATA | NEWS & EVENTS

SEARCH OUR SITE GO

THE DATA LIFECYCLE

We manage the research data lifecycle using innovation and technology to ensure an ongoing process of data creation, curation and use

READ ON

RE-USE CREATE SHARE

FIRST TIME HERE?
HELPFUL INFORMATION

A QUICK GUIDE TO THE ARCHIVE
2 of 10: We hold thousands of data collections for social science research and teaching, quantitative and qualitative

WHO GIVES US DATA?

What's new

DEPOSITING

Depositing your da they will be profes

DEPOSIT DATA

OUR DATA IN U

Labour Force Surv the health service resources equitab

OUR DATA IN USE

CONSENT & E

Dryad Digital Repository

www.datadryad.org

DRYAD

Login

Search Data

Submit Data Now!

See how to submit

Dryad is an international repository of data underlying peer-reviewed articles in the basic and applied biosciences. Dryad enables scientists to validate published findings, explore new analysis methodologies, repurpose data for research questions unanticipated by the original authors, and perform synthetic studies. Dryad is governed by a consortium of journals that collaboratively promote data archiving and ensure the sustainability of the repository.

As of Jun 2, 2012, Dryad contains **1668 data packages** and **4204 data files**, associated with articles in **131 journals**.

Recent Posts from the Dryad Blog

- Fossil preservation
- Chasing clinical trial data
- NSF provides further support to Dryad

My Account

Login or Register

Browse

Authors
Journal Title

Information

Depositing Data
Using Data
Dryad Partners
Journal Archiving Policy
About Dryad
Dryad Blog
Dryad Documentation

Recently Published Data

Deline B, Ausich W, Brett C (2012) Data from: Comparing taxonomic and geographic scales in the morphologic disparity of Ordovician through Early Silurian Laurentian Crinoids. *Paleobiology* doi:10.5061/dryad.c919g

Jasmin J, Zeyl C (2012) Data from: Life-history evolution and density-dependent growth in experimental populations of yeast. *Evolution* doi:10.5061/dryad.4142d

Evans M, Bernatchez L (2012) Data from: Oxidative phosphorylation gene transcription in whitefish species pairs reveals patterns of parallel and non-parallel physiological divergence. *Journal of Evolutionary Biology* doi:10.5061/dryad.s4kc8

Slater GJ, Harmon L, Alfaro M (2012) Data from: Integrating fossils with molecular phylogenies improves inference of trait evolution. *Evolution* doi:10.5061/dryad.q96d7

Mallatt JM, Craig CW, Yoder MJ (2012) Data from: Nearly complete rRNA genes from 371 Animalia: updated structure-based alignment and phylogenetic analysis. *Molecular Phylogenetics and Evolution* doi:10.5061/dryad.1v62kr3q

Dai L, Vorselen D, Korolev KS, Gore J (2012) Data from: Generic indicators for loss of resilience before a tipping point leading to population collapse. *Science* doi:10.5061/dryad.p2481134

Ensminger AW, Yassin Y, Miron A, Isberg RR (2012) Data from: Experimental evolution of Legionella pneumophila in mouse macrophages leads to strains with altered determinants of environmental survival. *PLoS Pathogens* doi:10.5061/dryad.95mt02sb

South SH, Arnqvist G, Servedio MR (2012) Data from: Female preference for male courtship effort can drive the evolution of male mate choice. *Evolution* doi:10.5061/dryad.8sb02

Data Publisher for Earth & Environmenta...

data-archive.ac.uk

Not logged in (log in or sign up)

PANGAEA®

Data Publisher for Earth & Environmental Science

All | Water | Sediment | Ice | Atmosphere

Search

Help | Advanced Search | Preferences | more...

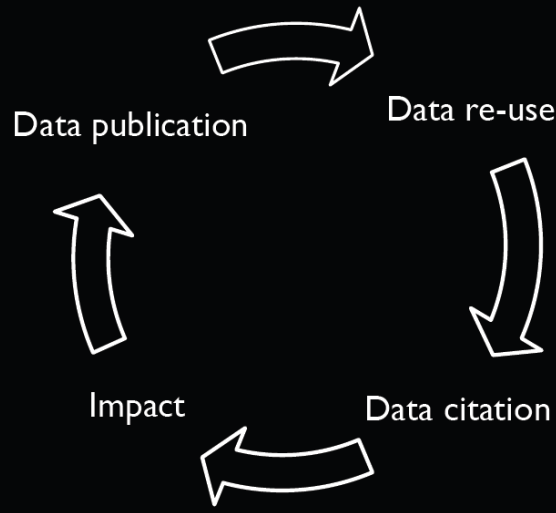
About – Submit Data – Projects – Software – WDC-MARE – Contact

This work is licensed under a Creative Commons License

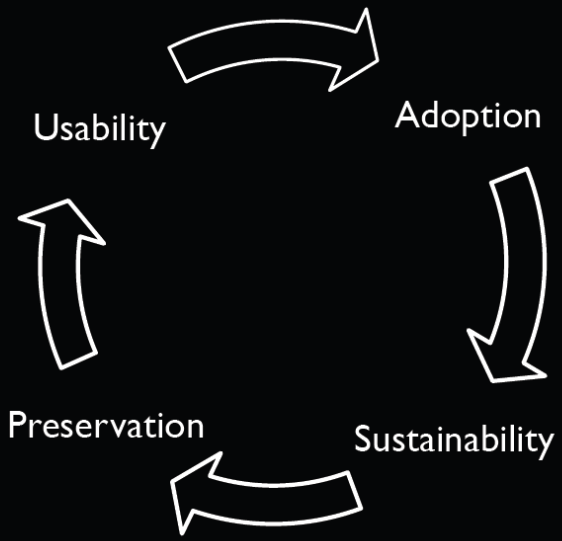
DPHEP

Sünje Dallmeier-Tiessen, Salvatore Mele
CERN – Scientific Information Service – Open Access Section
DPHEP@CHEP NYC May 24th 2012

AN OPPORTUNITY FOR HEP?



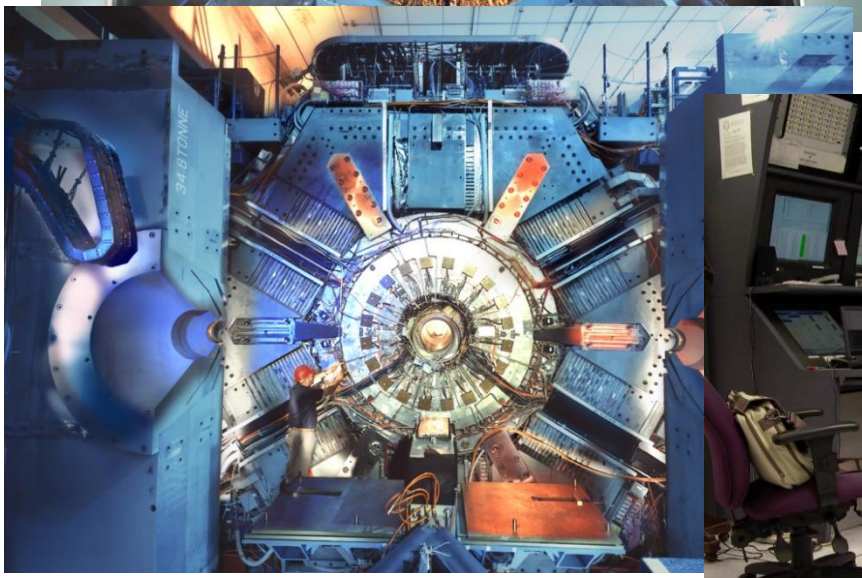
Virtuous circle of open data



Virtuous circle of preservation, re-use and (open) access

Sünje Dallmeier-Tiessen, Salvatore Mele
CERN – Scientific Information Service – Open Access Section
DPHEP@CHEP NYC May 24th 2012

THANK YOU!

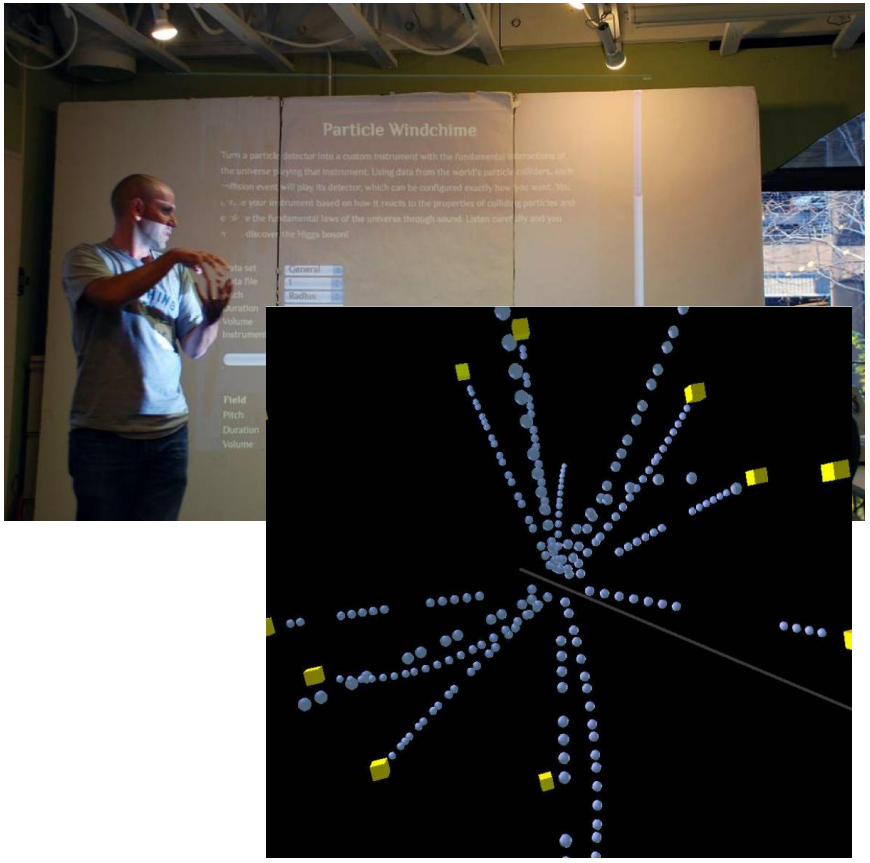


Backup

TO KNOW MORE…

OUTREACH: WIND CHIME

- Publicly accessible HEP data's impact was seen in Science Hack Day (SHD) events. One of the event's organizers, Ariel Waldman stated, "The mission of Science Hack Day is to get excited and make things with science! A Hack Day is a 48-hour-all-night event that brings together designers, developers, scientists and other geeks in the same physical space for a brief but intense period of collaboration, hacking, and building 'cool stuff'" (http://sciencehackday.com/). In 2010, Monte Carlo data from the BaBar experiment was brought to he SHD San Francisco event to provide the seed for a science/art mashup. Participants used the data to produce a Particle Physics Wind Chime website, that allowed the user to map detector and particle properties (momentum, detector hits, particle-ID, etc.) onto sonic characteristics (volume, pitch, timbre, etc.). The excitement of the participants was palpable as they "heard" the sounds of particle physics! One could even hear the differences between different physics events. This experience was featured in a BBC science podcast and in Symmetry Magazine. This was just featured in a panel on science outreach on the internet "Get Excited and Make Things with Science" in the past South by South West conference (http://schedule.sxsw.com/2012/events/event_IAP10977/).

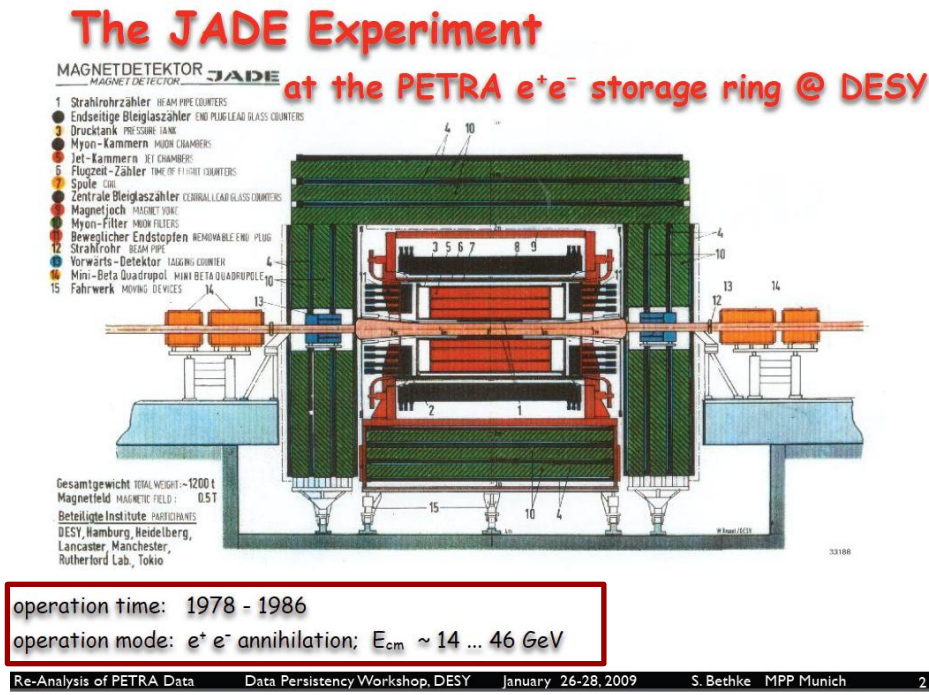


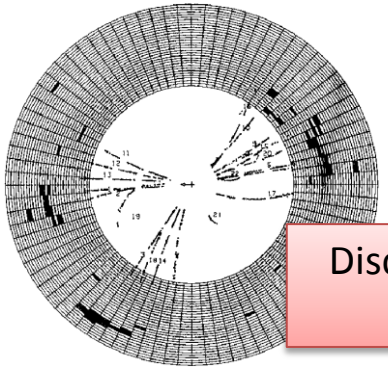
The BaBarian Matt Bellis

THE JADE ADVENTURE (I)

A STORY TOLD BY S. BETHKE @ DPHEP 1ST WORKSHOP

- Benefits of old data
  - re-do previous measurements:
    - Increased precision
    - Reduced systematics
  - perform new measurements:
    - at Energies and processes where no other data are available today (and in future)
  - if new phenomena found today:
    - go back and check at lower E



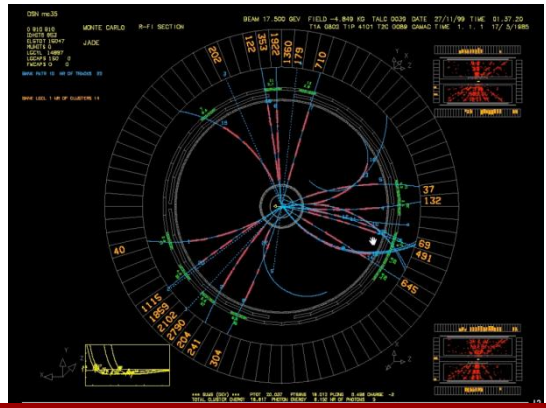


Discovery of the gluon in e+e- → qqbar g

THE JADE ADVENTURE (II)

A STORY TOLD BY S. BETHKE @ DPHEP 1ST WORKSHOP

- 1995: "private" initiatives *to :*
  - Rescue data from original archive tapes and copy them onto more modern media (IBM cartridges & Exabyte) (J. Olsson @ DESY)
  - Reanalyse data using modern (LEP-like) methods and observables plus improved theoretical calculations (S. B. and P. Movilla-Fernandez @ RWTH Aachen)
  - Revitalise JADE software on modern computer platforms to enable generation of new MC data files (P. Movilla Fernandez, J. Olsson)
  - Core software archived and saved at DESY, but routines kept on private user accounts were inevitably lost when removing and destroying old archive tapes. (e.g. JADE muon system libraries gone forever ... !)
- Since 1996, O(10) publications, O(10) conf. contributions;
- No competition in e^+e^- data analysis at E_{cm} ~14 ... 200 GeV

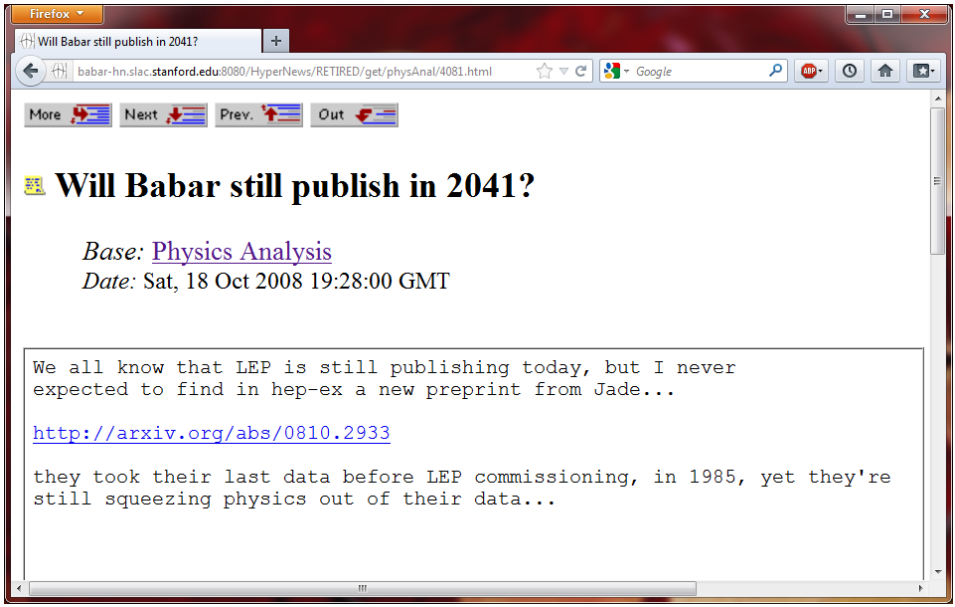


New Jade events display in colors!

THE JADE ADVENTURE (III)

A STORY TOLD BY S. BETHKE @ DPHEP 1ST WORKSHOP

- One important calibration file, containing the recorded luminosities of each run and fill, was stored on a private account and therefore lost when DESY archive was cleaned up.
  - Jan Olsson, when cleaning up his office in ~1997, found an old ASCII-printout of the JADE luminosity file. Unfortunately, it was printed on green recycling paper - not suitable for scanning and OCR-ing.
  - A secretary at Aachen re-typed it within 4 weeks. A checksum routine found (and recovered) only 4 typos.
- An old version of the original BOSlib 1979 version was found, on our request, at the Tokyo computer centre.
- Peter Bock, when cleaning out an old lab at the Physics Institute at Heidelberg University, found a few 9-track tapes containing original JADE MC files which were very valuable for validating results of our first re-analyses in ~1997



BLIND ANALYSIS

- BaBar has adopted the use of *blind analysis* whenever possible to avoid the analyst's bias when determining the analysis strategy (the selection cuts, for example).
  - The blinding is done in several ways. If the signal region is known then the analyst never looks at the signal box or, for measurements of parameters, the answer is hidden by an unknown offset.
  - More difficult, but still doable, is dealing with measures for which the study of the signal is fundamental or, even worse, when going *bump-hunting.*
- A BaBar analysis must be practically complete and gone through several layers of internal review before the analysts are allowed to unblind the result.

Firefox

Initial-State Radiation Measurement of t...

inspirehep.net/record/1086164?ln=en

Google

INSPIRE HEP

Welcome to INSPIRE! INSPIRE is now in full operation and supersedes SPIRES. SPIRES front end at all mirrors has been switched off. Please direct questions, comments or concerns to feedback@inspirehep.net.

HEP :: HEPNAMES :: INSTITUTIONS :: CONFERENCES :: JOBS :: EXPERIMENTS :: HELP

Information | References (48) | Citations (0) | Files | Plots

Initial-State Radiation Measurement of the $e^+e^- -> \pi^+\pi^-\pi^+\pi^-$ Cross Section.

BaBar Collaboration (J.P. Lees *et al.*) Show all 387 authors.

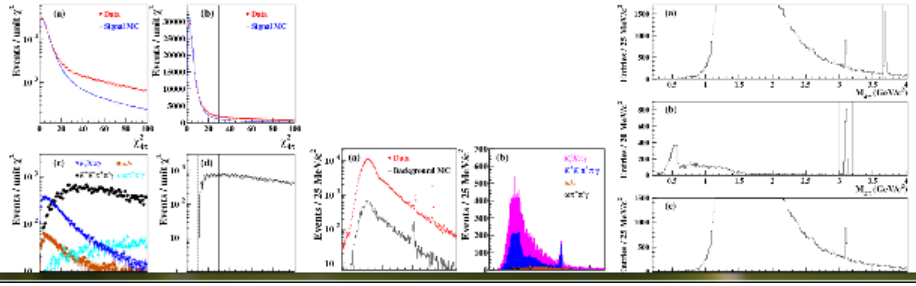
Jan 2012
19 pp.

Experiment: SLAC-PEP2-BABAR
SLAC-PUB-14857,BABAR-PUB-11-016
e-Print: **arXiv:1201.5677 [hep-ex]**

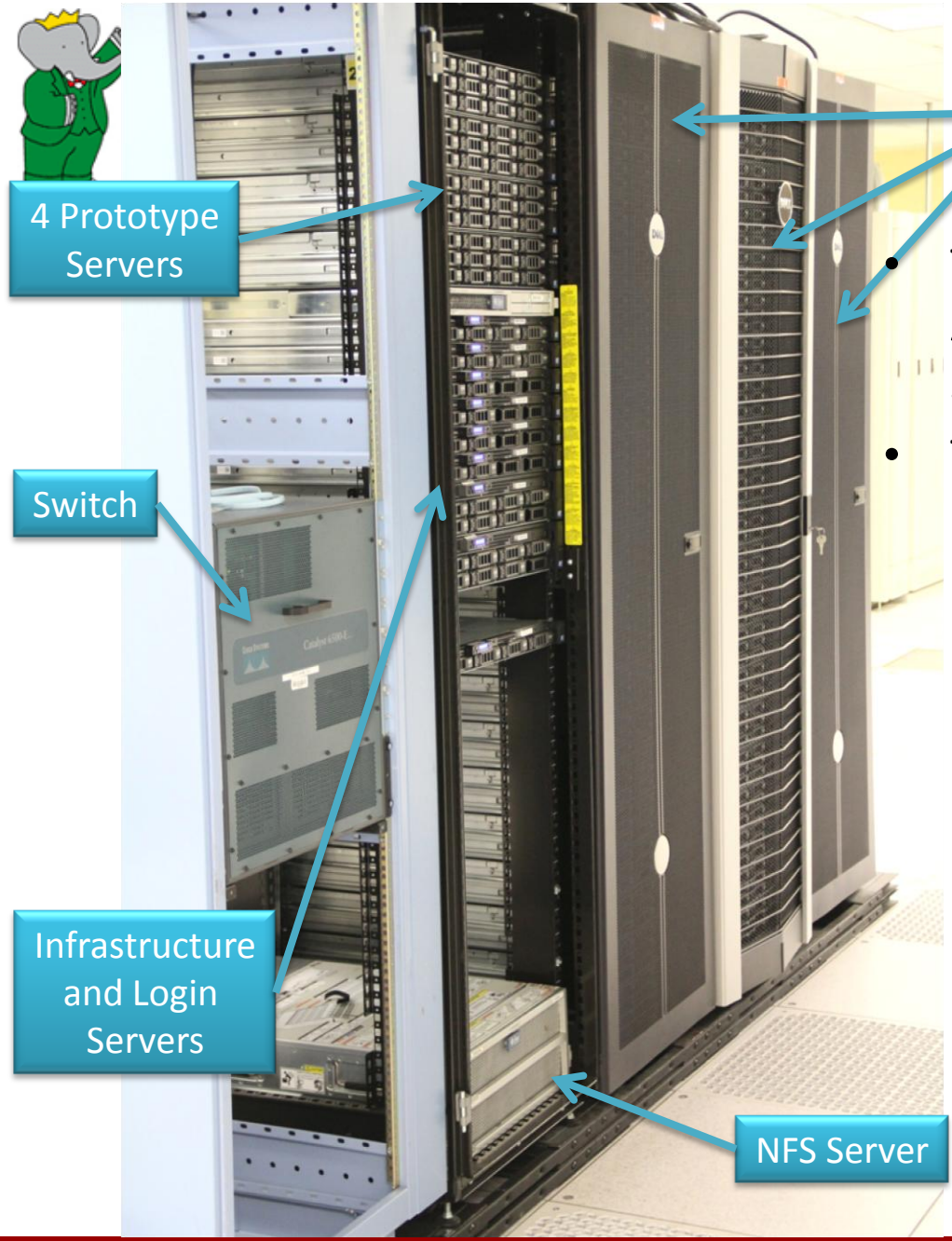
Abstract: We study the process e+e- -> pi+pi-pi+pi-gamma, with a photon emitted from the initial-state electron or positron, using 454.3 fb^-1 of data collected with the BABAR detector at SLAC, corresponding to approximately 260,000 signal events. We use these data to extract the non-radiative sigma(e+e- ->pi+pi-pi+pi-) cross section in the energy range from 0.6 to 4.5 Gev. The total uncertainty of the cross section measurement in the peak region is less than 3%, higher in precision than the corresponding results obtained from energy scan data.

Note: Long author list - awaiting processing

Keyword(s): INSPIRE: photon: emission | radiation: initial-state interaction | pi pi | initial state



Extended documentation. Inspire can also host password protected collaborations internal documents.



- The BaBar LTDA (Long Term Data Access) Cluster in its final setup in the SLAC computing building.
- The first example of Level4 preservation

H1, HERMES, ZEUS

Data Analysis Models @ DESY

- H1
  - preservation level 4
  - full chain from compilation of simulation, reconstruction and analysis code
  - full flexibility in the future for data and MC
- HERMES
  - preservation level 4
  - ADAMO-based micro-DST files for data and MC
- ZEUS
  - preservation level between 3&4
  - data and MC preserved in form of ROOT-based Common Ntuples
  - in addition maintain the ability of simulation of small samples of new MC in the future using existing executables

ALICE

CONCLUSIONS

- ALICE recognizes the importance of Data Preservation
- ALICE will work with the other experiments in order to contribute to and implement a common solution for DP
- ALICE would agree to Level 1 and Level 2
  - Provided resources are found!
- Level 3 seems harder and substantial resources should be found inside the experiment
  - It is however provided for members of the Collaboration
- Level 4 seems to be out of scope for the moment, unless there is a common decision by all the LHC experiments to do it
  - It is also provided for members of the Collaboration
- In any case computational resources will be needed to test the system (disk and CPU)

DATA PRESERVATION IN ALICE

FEDERICO CARMINATI

ATLAS DP Practical Steps

- Making sure raw data can be reprocessed long-term (Level 4)
  - Identifying key datasets for 'unique data' preservation
  - Setting up regular reprocessing and validation
- Ensuring the capability to run old trigger selections offline
- AODfixing may be useful
  - This means level 4 operations can be applied to level 3 AOD format
- Level 1 data
  - Extensive use of Inspire/HEPdata for level 1

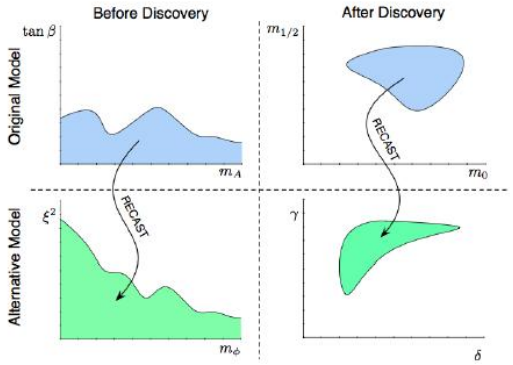
ATLAS Data Preservation and Access

Roger Jones

Practical steps - RECAST

arXiv:1010.2506

- Framework developed to extend impact of existing analyses
- Candidate for within-experiment and long-term analysis archival, encapsulating the full trigger & event selection, data, backgrounds, systematics
- Allow an existing analysis to be reinterpreted under an alternate model hypothesis
  - Complete information from original analysis, including the tacit information, contained in the data
  - Not optimized for the new model, but more reliable than a naïve reanalysis?



Recast seen as a very promising solution for preserving analyses and useful, cost effective preservation of information – addresses levels ~1-~3

CMS – DP Policy approved in March 2012

Guiding principles for the policy

- CMS is funded to fulfill its main goal → study physics.
- Therefore, for the most efficient use of the available resources, the following three principles were used as a guidance following in preparing the policy:

1) The greatest **benefit** from any data preservation activity should be **to the CMS collaboration** and collaborators in terms of scientific outcome, efficiency and preserved know-how.
2) All data preservation activities should be a **natural long-term extension** to the current CMS way of operation.
3) The data policy was required to include a protection of the collaboration from external use cases concerning open access that could generate burdens beyond the available funding and resources.

Kati Lassila-Perini on behalf of the CMS COllaboration 24.05.2012 4

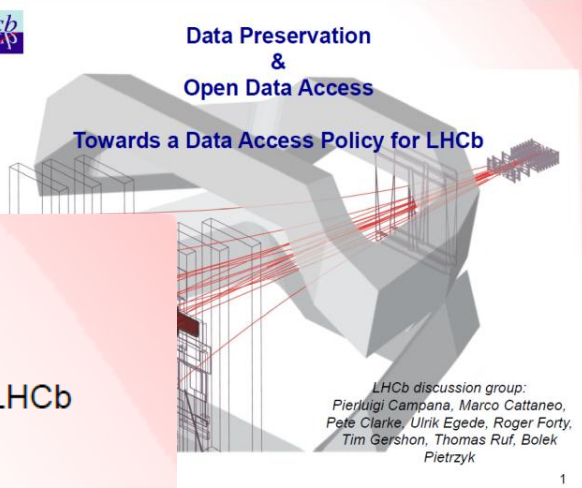
CMS (II)

The open access policy - extract

- At Level 1, the additional data are made available at the moment of the publication.
- At Level 2, simplified data format samples released promptly as determined by the CB.
- At Level 3 (reconstructed data), public data releases **yearly** accompanied by stable, open source, software and suitable documentation for analysis at Level 3
  - usually **three years after** data taking and during the long shut-downs, else at best effort
  - Collaboration Board can decide to release some particular data sets either earlier or later
  - **an upper limit to the amount of public data** 50% of the data available to CMS.
- The raw data formats of level 4 are not useful for analysis and will not be included in the public data release. Only level 3 formats after final calibration and reconstruction will be made available. These are the data formats used also internally in CMS for analysis.

Kati Lassila-Perini on behalf of the Data preservation task force 23.02.2012 17

LHCb



Draft Policy THIS IS NOT AGREED OR ENDORSED BY THE LHCB CB YET

1. Data preservation is fundamentally important for the collaboration itself <…….> LHCb will seek to develop such a data preservation capability as soon as practical.

2. LHCb supports the principle of open data access. We can envisage providing some such data access based upon the work needed internally for data preservation (point 1 above). However, as for other modern high-energy physics experiments, the data are complex, and making data available meaningfully requires substantial resources, therefore in the immediate future any provision in this sector will be modest.

3. Overall the collaboration expects to follow the policy developed by CERN and the LHC experiments jointly on these matters, after appropriate approval by the LHCb Collaboration Board.

4. LHCb is resource limited at present, and would therefore welcome the availability of additional resources from funders targeted at this important sector in order to realise aspirations of this document.

5. Access to its data by people outside the collaboration can be considered at four levels of increasing complexity, listed below, with associated conditions…….

20/05/2012 5

NATURAL ENVIRONMENT RESEARCH COUNCIL

All environmental data held by the NERC data centers will be made freely available without any restrictions on use […]. The policy introduces a formal requirement for all applications for NERC funding to include outline data management plans, which will be evaluated as part of the standard NERC grant assessment process.
http://www.nerc.ac.uk/research/sites/data/policy.asp

We believe that data sharing is essential for expedited translation of research results into knowledge, products, and procedures to improve human health. The NIH endorses the sharing of final research data to serve these and other important scientific goals. The NIH expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

NATIONAL INSTITUTES OF HEALTH

wellcome trust

The Wellcome Trust expects all of its funded researchers to maximise the availability of research data with as few restrictions as possible. […] In cases where the proposed research is likely to generate data outputs that will hold significant value as a resource for the wider research community, applicants will be required to submit a data management and sharing plan to the Wellcome Trust prior to an award bring made.
http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm

Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing. Proposals […] must include a supplementary […] "Data Management Plan" (DMP) […] describ[ing] how the proposal will conform to NSF policy on the dissemination and sharing of research results. http://www.nsf.gov/bfa/dias/policy/dmp.jsp

NSF

Science & Technology Facilities Council

Data resulting from publicly funded research should be made publicly available after a limited period, unless there are specific reasons (e.g. legislation, ethical, privacy, security) why this should not happen. The length of any proprietary period should be specified […] and justified, for example, by the reasonable needs of the research team to have a first opportunity to exploit the results of their research.
http://www.stfc.ac.uk/resources/pdf/stfc_scientific_data_policy.pdf

NASA

NASA has provided public insight into its operations for many years, from publishing its employee directory online to providing human capital information query-able in many ways. Since NASA's inception, we have publicly archived all of its data received from spacecraft projects, including over 4TB of new Earth Science data each day. There are tools and geodata catalogs available to allow scientists and the public to access NASA's raw data. http://www.nasa.gov/open/data.html

DPHEP

Sünje Dallmeier-Tiessen, Salvatore Mele
CERN – Scientific Information Service – Open Access Section
DPHEP@CHEP NYC May 24th 2012

SLAC NATIONAL ACCELERATOR LABORATORY