

A Web Based Document Database

*Computing in High Energy and Nuclear
Physics*

La Jolla, CA: 24–28 March, 2003

Eric W. Vaandering

ewv@fnal.gov

Vanderbilt University



Abstract

We will describe a document database developed for BTeV which has now been adopted for use by other collaborations within and without Fermilab. A single web based database and archival system is used to maintain public and private documents as well as documents for a related collaboration. The database performs archiving, versioning, access control, and easy remote access and submission. The talk will cover the technical and security requirements of the database and the implementation. Usage patterns, improvements in our collaborative style, and mis-steps along the way will also be discussed.



Outline

- Problems with our old solution
- Design Considerations
- Document Classification
- Implementation
- Effects on Collaboration
- Other Users and Lessons Learned
- Conclusions and URLs



BTeV

BTeV is an LHC era experiment scheduled to begin data taking in 2008. A reasonable expectation is that data taking will continue for about 5 years and that analysis may continue 5 or more years after that. So, we'd like a way of preserving our documentation until 2020 or so.

The system described here may not last that long, but it should be easily translated.



The Need for a New Database

BTeV had an old document catalog (saved URLs, not actual documents), but it was too limiting.

- Classified only by author-defined keywords.
- Updates replaced original.
- One author per document.
- One file per doc (two docs for PS and PDF).
- Separate lists for public and private documents (more duplication).
- From 1995–2001, created ~ 110 private, ~ 40 public docs. About 10 of these had disappeared.
- Drew some ideas from NuMI Notes.



Design of New Database

- All documents are kept on our web server, not scattered machines. No link rot.
- Each document can have multiple revisions and old revisions are still available.
- Each revision can have multiple files which accommodates different file types (source and presentable) and, *e.g.*, HTML trees.
- No limits on numbers of authors or topics.
- Document ID is just a number (no topics or public/private).
- A single database for talks from group meetings, conference documents, and publications *and* presents these special cases in special ways.



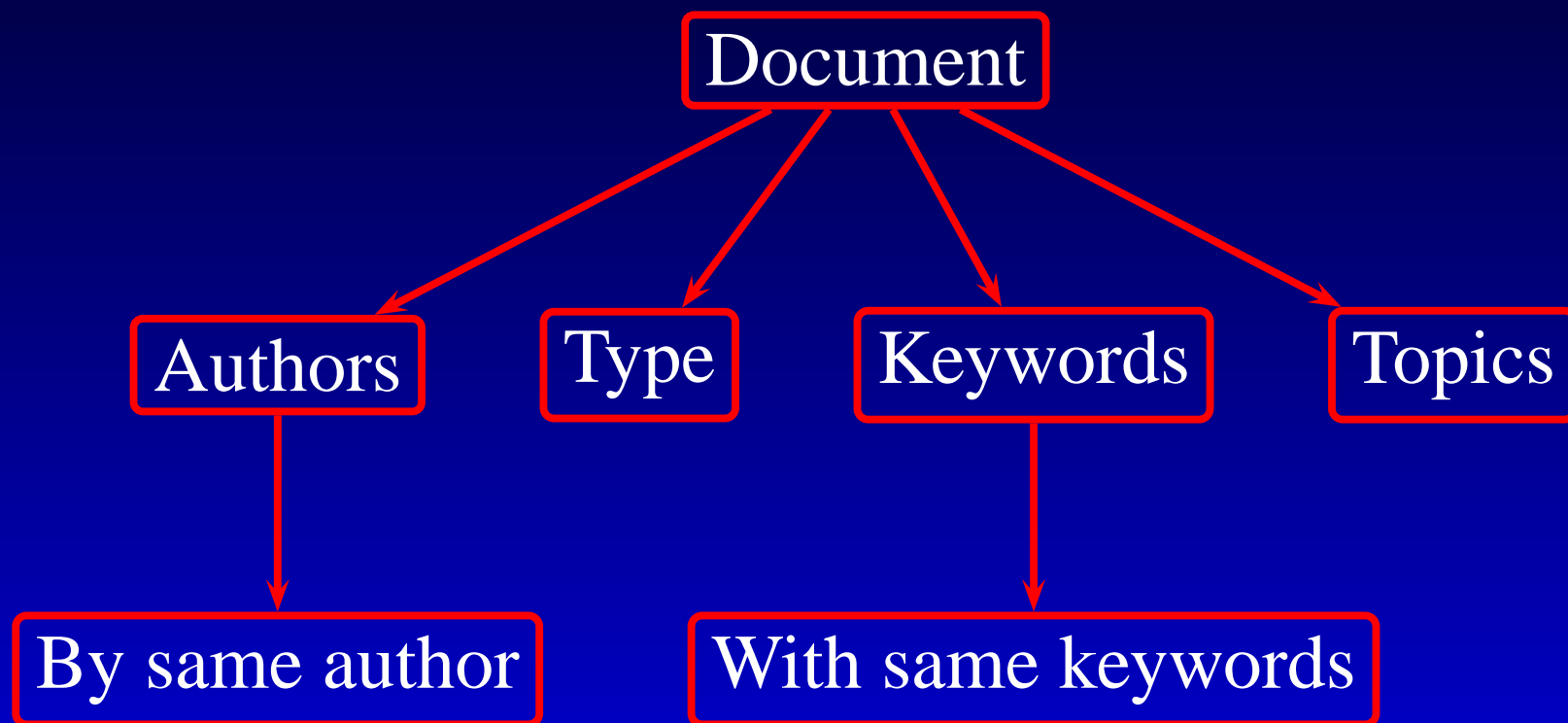
Security Considerations

- Reviewers need access to selected documents but cannot create or modify.
- BTeV has “associated groups” that need to create documents and view selected documents.
- Want documents that are only accessible to sub-groups like the executive council.
- Users only know that a document exists if they have privileges to view it.
- Wanted the ability to easily move documents, or just certain versions of documents, from restricted to public and vice-versa.
- Access control via standard HTTP authorization.



Document Classification

Each way of classifying a document is hyperlinked to other documents of the same kind.



Can quickly navigate, find documents of interest.



Other Document Information

There are several other pieces of information we store:

- Title and Abstract
- Creation and modification times
- Groups that may view/modify the document
 - As an option, these lists can be different.
- Hyperlinked references (journals, arXiv.org, etc.)
- Additional publication information

These and the classifications are all searchable.



Implementation

The system consists of three parts:

- Meta-data is stored in a MySQL database.
- The files in a document are served by a web server from the regular file system.
 - Contents of files are not directly searchable, may change by incorporating a search engine like **htdig**.
- A web interface for adding, changing, searching, and viewing documents (Perl/CGI).
 - Easy to use for non-technical users.
 - Flexibility to allow for changing, fixing mistakes.
 - Presentation is configurable to user preferences.



Web Interface

- For modifying documents, lots of options:
 - Reserve a slot for a future document
 - Create a new document
 - Create an updated version
 - Update meta-data for a document
 - Add or *replace* files in a document
 - *Can* circumvent versioning, haven't seen this.
- User has several ways of supplying files
 - Upload from the users computer
 - Fetch URLs (SSL or password protected too)
 - tar/zip files (good for photo collections, HTML, etc.)



List of Documents



Document List by Topic

Search

BTeV at Work

General

Computing

Documents

Physics

Detector

IR & Tevatron


These documents on **Collaboration Meetings:14-15 March 2003** are available:

[Agenda for this meeting](#)

Title	Author	Topic(s)	Files
BTeV RICH Status Report	Marina Artuso	Detectors:RICH	pdf file btev-mar03-reva.pdf
Open Plan Status - March 2003 Collaboration Meeting	E. Barsotti	BTeV General:WBS	Open Plan Status - March 2003 Collaboration Meeting PDF file Open Plan Status - March 2003 Collaboration Meeting PowerPoint file
Agenda-BTeV March, 2003 Collaboration Meeting	Joel N. Butler	None	agenda_mar03.html
Investigating Pythia Tunes	Lynn A. Garren	Computing:Simulation	index.html
Trigger Status	Erik E. Gottschalk	None	Powerpoint pdf - slides 9 & 11 didn't convert quite right.
Straw Status, March 2003	A. Hahn	Detectors:Straws	PPT file PDF file this one is gigantic! (jpegs seem to expand tremendously)
DAQ Status	K. Honscheid et. al.	None	Powerpoint PDF
BTeV Muon System Update	Will E. Johns	Detectors:Muon	PDF of talk powerpoint of talk
RTES Status 3/14/03	Jim Kowalkowski	Computing:RTES	RTES status 3 14 03
EMCAL Status	Y. Kubota	Detectors:EM Cal	talk in pdf talk in ppt
Pixel Detector Status Report	Simon W. L. Kwan	Detectors:Pixels	PDF file Power point file



Example Document View



BTev Document 140-v4

Search

BTev at Work

General

Computing

Documents

Physics

Detector

IR & Tevatron

BTev Document Database Design and Interface

Requested by: [Eric Vaandering](#)

Updated by: [Eric Vaandering](#)

Document #: BTev-doc-140-v4

Document type: [Note](#)

Document Created: 19 Dec 2001, 17:13

Contents Revised: 08 Jun 2002, 11:00

DB Info Revised: 08 Jun 2002, 11:00

Authors:

- [Lynn A. Garren](#)
- [Eric Vaandering](#)

Topics:

- [Computing:Web Pages & Tools](#)

Restricted to:

- BTev
- RTES

Abstract:

In this memo we describe the design and implementation of the BTev document database. The document database consists of three parts: a relational database which stores information about the documents, a file system hierarchy used for storing the documents themselves, and a suite of CGI scripts which provide coherent access to both sets of information.

Keywords:

[DocDB](#)

Files in Document:

- [PDF version](#) (doc_db.pdf, 124.1 kB)
- [Postscript version](#) (doc_db.ps, 124.8 kB)

Other Files:

- [TeX source](#) (doc_db.tex, 14.7 kB)
- [xfig source for figure](#) (DocDB_model.fig, 9.2 kB)

Retrieve archive of files as [tar.gz](#), [zip](#).

Update Document

Update DB Info

Add Files to Document

Other Versions of this document:

- [BTev-doc-140-v3](#) (11:19:48 01 Jun 2002)
- [BTev-doc-140-v2](#) (14:33:51 20 Dec 2001)
- [BTev-doc-140-v1](#) (17:55:38 19 Dec 2001)



Other Features

- Very flexible search capability.
- E-mail notifications allow users to receive e-mail on changed documents.
 - Triggered by topics/subtopics, authors, and keywords user is interested in.
 - Any combination of these for immediate, daily, or weekly notification.
- A web interface for database administrators:
 - Modify authors, topics, groups, etc.
 - Delete documents.
 - Never (directly) touch SQL database —a big plus.



Impact on Collaboration

- Redefining what a “document” is:
 - Not just words written on a piece of paper, but any information you want to save and share.
 - We’ve been somewhat effective with this.
- More information is easily available.
 - In past, sub-group web pages maintained documents.
 - With an easier to use and more visible system, this information is now being placed in the database.
 - Users no longer feel their document has to reach some level of “importance” before placing it in the database.
 - Migration of “legacy” documents is slow.



Collaboration Meetings

BTeV has a video conference about every 6 weeks, about 20 talks per meeting.

- In the past, a secretary collected URLs, maintained agenda.
 - Difficult for weekend meetings, time lag problems.
 - Result: A flurry of e-mails and confusion for late arriving documents.
- Now, the whole process is much smoother.
 - Updates, reactions happen minutes before presentation.
 - Specialized view and (optional) entry form for talks.
 - Talks are available from same place as other documents.



Reviews

BTeV is under active review. The DocDB has been a great help here too.

- We have a read-only account for reviewers.
- Relevant documents are made accessible to reviewers.
- Keywords used effectively to categorize documents for reviews.
- Our latest review required 250 documents.
 - Many will remain the same or be updated for future reviews.



Usage observations

- Old system, 1995–2001: About 150 documents.
- New system went live at end of 2001.
- Currently > 1600 documents:
 - Some 400+ of these are “legacy” documents.
 - Collaboration (~ 200 people) producing 3–4 docs/day.
- Statistics:
 - Average document: 1.6 versions
 - Each version: 1.5 authors, 1.8 topics, 1.8 files.
 - Total of almost 5000 files being tracked (5.2 GB).
- Number of “living documents” is small.



Other Users

Setting up this database takes a couple of hours, becoming easier.

- Fermilab Beams Division:
 - Began with Main Injector group, quickly adopted by division.
 - In active use for several months, now > 500 documents.
 - Seem to be more effective in enforcing organization.
 - Also helpful for reviews (less manual organization).
 - Previously lots of meetings were hard-copy only.
- US-CMS is evaluating.



What We've Learned

Overall this has been a very good experience, but you can always do better.

- Organization:
 - Concept of keywords was added later.
 - Some sub-groups have been very proactive in this regard, others less so.
 - Never had “official” guidance on use. This may be a good thing, may not be.
- Initially lacked some flexibility for fear of misuse.
- Wasn't initially designed to be portable, but this wasn't too hard to change.
- Obvious: Use a system like this early on.



Conclusions and URLs

- We have built and are successfully running a useful document database which has made real improvements to our collaborative style.
- It is being adopted by other groups; we're willing to help others as well.
- URLs

- BTeV Public Server:

<http://www-btev.fnal.gov/cgi-bin/public/DocDB/DocumentDatabase>

- Beams Division (lots of public documents):


<http://beamdocs.fnal.gov/cgi-bin/public/DocDB/DocumentDatabase>

- Download, instructions, demo server (soon):

<http://cepa.fnal.gov/DocDB/doc/install-docdb.html>



Entry Form



Document Addition

Search

BTeV at Work

IR & Tevatron

GeneralComputingDocumentsPhysicsDetector

You have chosen to add a new document to the database.
Make sure you know all the required information and the file is ready for uploading.
Click on any of the headings for help.

Title:

Abstract:

Keywords: (space separated)

Upload by HTTP:

URL:

Description: ☒ Main?

User: Password:

Make sure you supply a format everyone can view (e.g. PDF) along with source formats.

Document type:

☐ Audio ☐ Proceeding ☐ Note



Entry Form, Part 2

- ☐ Video ☐ Other ☐ Technical Drawing
☐ Talk ☐ Figure/Photo ☐ Minutes
☐ Publication ☐ Poster

Requester:

Madhura Tamhankar
 Zhijing Tang
 Alexey Titov
 Terry Tope
 P. Torre
 Alexandre Toukhtarov
 Marcos Turqueti
 Lorenzo Uplegger
 A. Uzunian
 Eric Vaandering

Authors:

Madhura Tamhankar
 Zhijing Tang
 Alexey Titov
 Terry Tope
 P. Torre
 Alexandre Toukhtarov
 Marcos Turqueti
 Lorenzo Uplegger
 A. Uzunian
 Eric Vaandering

Security:

Public
 BTeV
 Executive
 RTES
 DocDBAdm
 Review

Topics:

BTeV General

Budgets
 ES&H
 Education & Outreach
 Electrical
 General
 Mechanical
 Organization
 Policies

Collaboration Meetings

14-15 March 2003
 31 January 2003
 6-7 December 2002
 12-13 October 2002
 30-31 August 2002
 17-18 July 2002
 31 May 2002
 5-6 April 2002

Computing

Analysis Software
 Control & Monitoring
 DAQ & Readout
 Data Archival
 Data QA
 Farms
 Framework
 Grid

Conferences

CHEP03
 NSS2002
 VERTEX2002
 Pixel 2002
 2002 IEEE NSREC
 Beach2002
 Beauty '02
 dpf2002

Detectors

Beams
 CZero
 EM Cal
 General
 Magnets
 Muon
 Other
 Pixels

Other Meetings

Software workshop, Fall 2001
 RTES - February 2002
 RTES - May 2002
 Integration Workshop - May 2002
 RTES - March 2002
 RTES - August 2001
 RTES - November 2001
 RTES - March 2003

Physics-Other

Beyond SM
 Electroweak
 Production
 QCD
 Rare Decays
 Standard Decays

Physics-b

Baryons
 Beyond SM
 CP Violation
 General
 Mixing
 Other
 Production
 Rare B decays

Physics-c

