A new Information Architecture, Website and Services for the CMS Experiment

Lucas Taylor
Head of Communications, CMS Collaboration at CERN
Fermi National Accelerator Laboratory, Batavia, Illinois, USA
E-mail: Lucas.Taylor@cern.ch

Eleanor Rusack
Fermi National Accelerator Laboratory, Batavia, Illinois, USA

Vidmantas Zemleris
Vilnius University, Lithuania and École Polytechnique Fédérale de Lausanne, Switzerland

Abstract. The age and size of the CMS collaboration at the LHC means it now has many hundreds of inhomogeneous web sites and services, and hundreds of thousands of documents. We describe a major initiative to create a single coherent CMS internal and public web site. This uses the Drupal web Content Management System (now supported by CERN/IT) on top of a standard LAMP stack (Linux, Apache, MySQL, and php/perl). The new navigation, content and search services are coherently integrated with numerous existing CERN services (CDS, EDMS, Indico, phonebook, Twiki) as well as many CMS internal Web services. We describe the information architecture; the system design, implementation and monitoring; the document and content database; security aspects; and our deployment strategy, which ensured continual smooth operation of all systems at all times.

1. Introduction

The Compact Muon Solenoid (CMS) collaboration at CERN's Large Hadron Collider (LHC) involves 3275 physicists and 790 technical staff from 179 institutes in 41 countries. These people are organised in more than 60 sub-groups and projects. Together they have produced more than 700,000 documents over two decades describing all aspects of the CMS design, construction, operations, performance and analysis. These range from published papers, reports, and notes to less formal working documents such as presentations, minutes, technical drawings, data sheets, and photographs.

These documents are strategic to the success of the multi-decade CMS research programme, therefore they must be stored securely and be readily accessible to CMS scientists. In 2010 a review of the situation found that, while more than 200,000 documents were already stored in managed document systems, about 400,000 more were not. Instead, they resided on about 1000 different websites, on the CERN afs and dfs file systems, and on private users' computers. In general such documents were not indexed, so finding them was next to impossible, and many were at risk of being lost forever. The

CMS Communications Group therefore developed a programme to consolidate the CMS knowledge base, with two main goals:

- **Goal 1: Manage all CMS documents in a document database; and**
- **Goal 2: Create a coherent web interface to all CMS information.**

In order to achieve these two goals, the new CMS information architecture had to address the issues of document *organization, labelling, search* and *navigation.* The strategy envisaged using a number of "Document Database" systems and a web "Content Management System".

Prior to embarking on a major technical programme to upgrade the CMS information systems, we first considered the potential benefits compared to the expected costs.

2. Do the benefits justify the costs?

The costs comprise support for three part-time people (the authors), summing to one full-time equivalent person (1 FTE). They have various skills in management, web systems and administration.

The primary benefit is to secure for many decades the huge amount of information (Figure 1.) that is critical to CMS operations, maintenance, upgrades, analysis and eventual decommissioning. This includes technical drawings and specifications, photographs, procedures, analysis techniques and results, and many miscellaneous items. This benefit is hard to quantify, but the loss of such critical information could clearly have serious technical and cost implications.

Secondary benefits come from improving CMS productivity. Making it quicker to find the correct information yields immediate efficiency gains. CMS meetings – 15,000 per year with 50,000 documents – tie up an integrated effort of more than 100 FTE continuously, so even modest efficiency gains are valuable. The running costs of the LHC, the CMS detector and the grid are large and the CMS fraction is very roughly 10^8 CHF per year for 10^7 seconds of data taking [1]. Even modest efficiency gains above the current 92% data-taking efficiency have a large payback.

A printed pile of all CMS documents that are already in a managed system
= 1.0 x (Empire State building)
Plus we have almost the same amount spread all over the place (PCs, afs, dfs, various websites …)

Figure 1. As of 2012, a stack of all the CMS documents in managed document systems would be the height of the Empire State building. A similar number are spread over "unmanaged" systems.

For each of these cases, the benefit could easily exceed the current 1 FTE cost, for example:

- 8 FTE saved if each CMS person wastes 1 minute less per day searching for information.
- 15 FTE saved if better information flow results in a 10% a reduction in the number, length or attendance of meetings (e.g. goals defined, documents posted in advance, minutes, etc.).
- 1 MCHF gained if data-taking efficiency increases by one percentage point.

Such savings are not unreasonable targets and further productivity gains could be expected in other areas, such as training, user support, grid operations, or data analysis. Even if all the savings cannot be fully realised, the total benefits still outweigh the very modest costs and even suggest that more investment in improving the CMS information systems would be well justified.

3. Achieving Goal 1: "Manage all CMS Documents in a Document Database"

3.1. Document Management Systems

A number of document management systems from CMS, CERN and Fermilab help CMS to address the challenges of organisation, labelling, search and navigation of information. CMS uses a number of document management systems from CMS, CERN and Fermilab (Table 1). Differences between these systems arise due to valid users' needs, so we cannot standardise on a single system but must work with a number of systems. All such systems have a number of common elements, as follows.

Documents is the generic name for all types of information content, including web pages, publications, notes, presentations, plots, technical drawings, data sheets, photos and so on. A document may correspond to a single file or may contain multiple files. *Document metadata* includes, for example, the document type, title, authors, date, description, topics or keywords, and so on. It enables documents to be classified in a variety of ways for subsequent browsing and searching. The *storage system* includes a file storage system for the content itself, for example a Unix server, and a relational database system for the metadata.

A web *user interface* is the portal to the underlying storage systems. It enables management functions such as the upload of documents and metadata and the creation of new content such as web pages. Crucially, the existence of metadata for each document enables effective interfaces to be created to browse and search for documents. This is particularly needed for non-text documents such as images.

Table 1. The main document database management systems used by CMS.

Document Management System		**Main CMS usage**
CADI	CMS Analysis Database Interface, from CMS	Manage workflows for preparing CMS papers and notes; once completed they are entered into CDS.
CDS	CERN Document Server	Official CMS papers, reports, notes and multimedia.
CINCO	CMS conference system	Manage workflows for abstract preparation, speaker selection, approval of talks and write-ups.
DocDB	Fermilab Document Database [2]	User-friendly system for any documents that do not naturally belong to another document system.
EDMS	CERN Engineering and Equipment Data Management System	Engineering drawings and associated technical data.
Indico	CERN Integrated Digital Conference system	Manage meeting agendas, presentations, minutes and related documents.
Twiki	CERN Wiki collaborative tool	Web pages and attachments, with many documents for CMS operations and analysis activities.

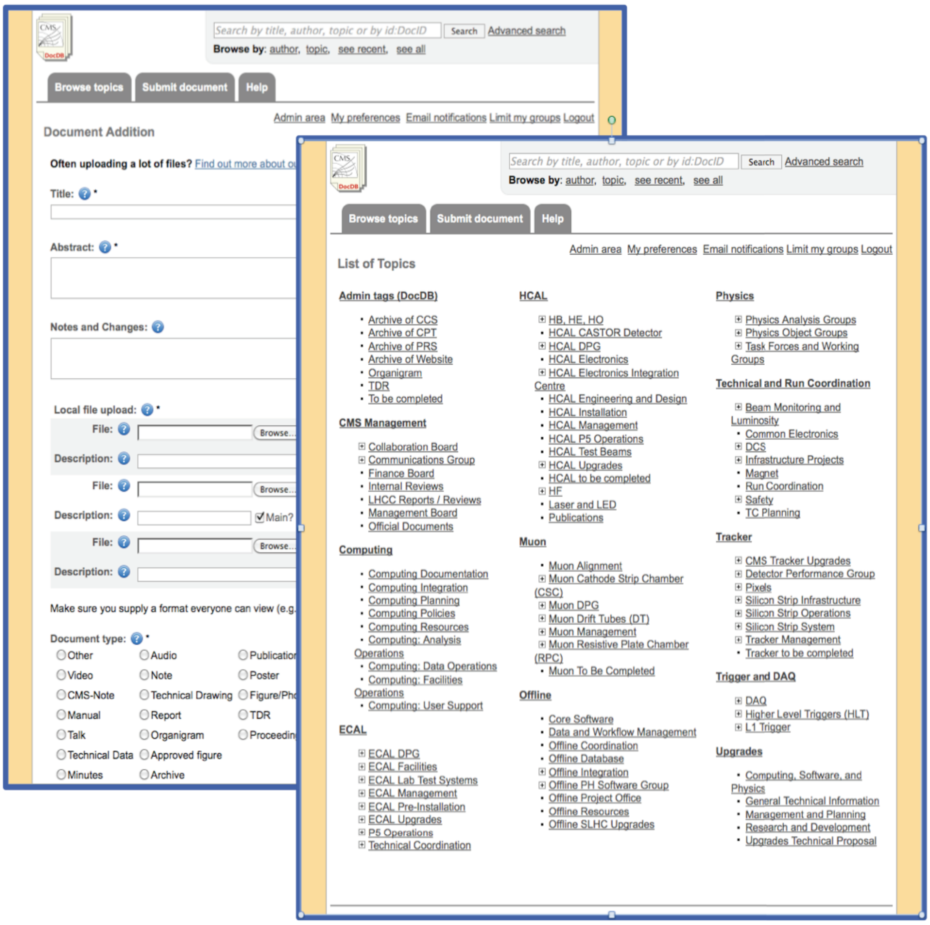
3.2. Tools to find CMS documents and import them into a document database

In January 2010, about 400,000 "unmanaged" documents resided outside any document management system. About 200,000 of these were found on CMS afs group space and collaboration websites. A further 100,000 reside on users' disk servers and personal computers (estimated by extrapolating from a few users who systematically searched their systems). Finally, the CMS internal and public wikis contained about 100,000 document files, 20% of which are web pages and 80% of which are attached files such as diagrams, plots and data files.

It was decided to process these documents and decide which to import into the DocDB database, together with associated metadata. DocDB has a simple interface (see Figure 2), a well structured set of topics, and already contains all possible CMS authors, 970 of whom have one or more documents in DocDB.

Recognising that it was unrealistic to expect users to process many thousands of documents individually, we developed a *Bulk Uploader* tool to facilitate the process. Using bash and Python scripts and a django-based web interface it supports the following steps:

1. Search file systems for documents (files), filtering by type and location, and organise them in batches, depending on origin, topic, user, etc. Web crawlers are used to search websites.
2. Automatically extract metadata from document files (e.g. title, author, etc.) and create thumbnail images, where possible.
3. The web interface, shown in Figure 3, permits users to examine the contents of their batches, select which documents to upload to DocDB, add or change the metadata, and schedule their documents for upload either individually or in groups.
4. Perform the bulk import of documents and metadata into DocDB.



Document Addition

Often uploading a lot of files? Find out more about o

Title: *

Abstract: *

Notes and Changes:

Local file upload: *

File: Browse...

Description:

File: Browse...

Description: ☑ Main?

File: Browse...

Description:

Make sure you supply a format everyone can view (e.g.

Document type: *

○ Other	○ Audio	○ Publicatio
○ Video	○ Note	○ Poster
○ CMS-Note	○ Technical Drawing	○ Figure/Ph
○ Manual	○ Report	○ TDR
○ Talk	○ Organigram	○ Proceedin
○ Technical Data	○ Approved figure	
○ Minutes	○ Archive	

List of Topics

Admin tags (DocDB)

- Archive of CCS
- Archive of CPT
- Archive of PRS
- Archive of Website
- Organigram
- TDR
- To be completed

CMS Management

- ⊞ Collaboration Board
- ⊞ Communications Group
- Finance Board
- Internal Reviews
- LHCC Reports / Reviews
- Management Board
- Official Documents

Computing

- Computing Documentation
- Computing Integration
- Computing Planning
- Computing Policies
- Computing Resources
- Computing: Analysis Operations
- Computing: Data Operations
- Computing: Facilities Operations
- Computing: User Support

ECAL

- ⊞ ECAL DPG
- ⊞ ECAL Facilities
- ⊞ ECAL Lab Test Systems
- ⊞ ECAL Management
- ⊞ ECAL Pre-Installation
- ⊞ ECAL Upgrades
- ⊞ P5 Operations
- ⊞ Technical Coordination

HCAL

- ⊞ HB, HE, HO
- HCAL CASTOR Detector
- ⊞ HCAL DPG
- HCAL Electronics
- ⊞ HCAL Electronics Integration Centre
- HCAL Engineering and Design
- HCAL Installation
- HCAL Management
- HCAL P5 Operations
- HCAL Test Beams
- ⊞ HCAL Upgrades
- HCAL to be completed
- ⊞ HF
- Laser and LED
- Publications

Muon

- Muon Alignment
- ⊞ Muon Cathode Strip Chamber (CSC)
- ⊞ Muon DPG
- ⊞ Muon Drift Tubes (DT)
- ⊞ Muon Management
- ⊞ Muon Resistive Plate Chamber (RPC)
- Muon To Be Completed

Offline

- Core Software
- Data and Workflow Management
- Offline Coordination
- Offline Database
- ⊞ Offline Integration
- ⊞ Offline PH Software Group
- Offline Project Office
- Offline Resources
- Offline SLHC Upgrades

Physics

- ⊞ Physics Analysis Groups
- ⊞ Physics Object Groups
- ⊞ Task Forces and Working Groups

Technical and Run Coordination

- ⊞ Beam Monitoring and Luminosity
- Common Electronics
- ⊞ DCS
- ⊞ Infrastructure Projects
- Magnet
- Run Coordination
- ⊞ Safety
- TC Planning

Tracker

- ⊞ CMS Tracker Upgrades
- ⊞ Detector Performance Group
- ⊞ Pixels
- ⊞ Silicon Strip Infrastructure
- ⊞ Silicon Strip Operations
- ⊞ Silicon Strip System
- ⊞ Tracker Management
- Tracker to be completed

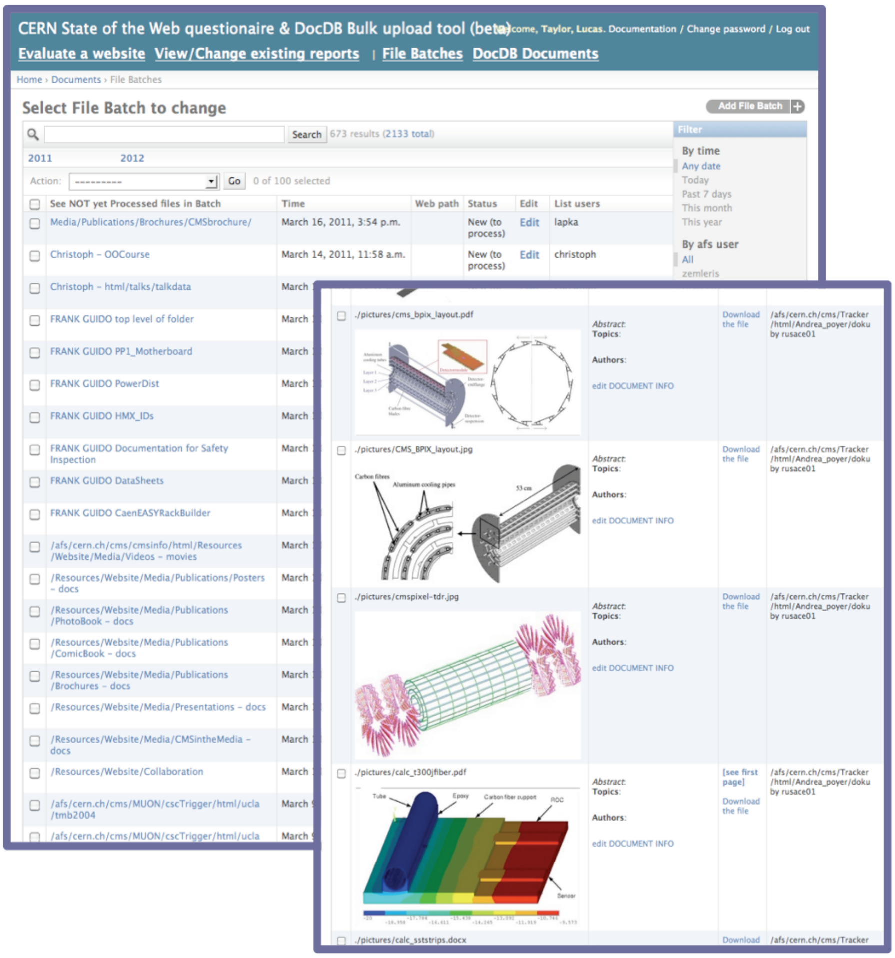
Trigger and DAQ

- ⊞ DAQ
- ⊞ Higher Level Triggers (HLT)
- ⊞ L1 Trigger

Upgrades

- Computing, Software, and Physics
- General Technical Information
- Management and Planning
- Research and Development
- Upgrades Technical Proposal

Figure 2. Web interfaces for uploading (left) and browsing (right) the CMS instance of the DocDB document database system.



CERN State of the Web questionaire & DocDB Bulk upload tool (beta) Welcome, Taylor, Lucas. Documentation / Change password / Log out

Evaluate a website View/Change existing reports | File Batches DocDB Documents

Home › Documents › File Batches

Select File Batch to change

Add File Batch +

Search 673 results (2133 total)

2011 2012

Action: --------- Go 0 of 100 selected

□	See NOT yet Processed files in Batch	Time	Web path	Status	Edit	List users
□	Media/Publications/Brochures/CMSbrochure/	March 16, 2011, 3:54 p.m.		New (to process)	Edit	lapka
□	Christoph – OOCourse	March 14, 2011, 11:58 a.m.		New (to process)	Edit	christoph
□	Christoph – html/talks/talkdata	March				
□	FRANK GUIDO top level of folder	March				
□	FRANK GUIDO PP1_Motherboard	March				
□	FRANK GUIDO PowerDist	March				
□	FRANK GUIDO HMX_IDs	March				
□	FRANK GUIDO Documentation for Safety Inspection	March				
□	FRANK GUIDO DataSheets	March				
□	FRANK GUIDO CaenEASYRackBuilder	March				
□	/afs/cern.ch/cms/cmsinfo/html/Resources /Website/Media/Videos – movies	March				
□	/Resources/Website/Media/Publications/Posters – docs	March				
□	/Resources/Website/Media/Publications /PhotoBook – docs	March				
□	/Resources/Website/Media/Publications /ComicBook – docs	March				
□	/Resources/Website/Media/Publications /Brochures – docs	March				
□	/Resources/Website/Media/Presentations – docs	March				
□	/Resources/Website/Media/CMSintheMedia – docs	March				
□	/Resources/Website/Collaboration	March				
□	/afs/cern.ch/cms/MUON/cscTrigger/html/ucla /tmb2004	March				
□	/afs/cern.ch/cms/MUON/cscTrigger/html/ucla	March				

Filter

By time
Any date
Today
Past 7 days
This month
This year

By afs user
All
zemleris

□	./pictures/cms_bpix_layout.pdf	*Abstract*: **Topics**: **Authors**: edit DOCUMENT INFO	Download the file	/afs/cern.ch/cms/Tracker /html/Andrea_poyer/doku by rusace01
□	./pictures/CMS_BPIX_layout.jpg	*Abstract*: **Topics**: **Authors**: edit DOCUMENT INFO	Download the file	/afs/cern.ch/cms/Tracker /html/Andrea_poyer/doku by rusace01
□	./pictures/cmspixel-tdr.jpg	*Abstract*: **Topics**: **Authors**: edit DOCUMENT INFO	Download the file	/afs/cern.ch/cms/Tracker /html/Andrea_poyer/doku by rusace01
□	./pictures/calc_t300jfiber.pdf	*Abstract*: **Topics**: **Authors**: edit DOCUMENT INFO	[see first page] Download the file	/afs/cern.ch/cms/Tracker /html/Andrea_poyer/doku by rusace01
□	./pictures/calc_sststrips.docx		Download	/afs/cern.ch/cms/Tracker

Figure 3. Web interface of the CMS bulk document upload tool.

Users of the bulk uploader can easily upload hundreds of documents per day into DocDB, which is much faster than using the standard DocDB interface. Between 2003 and 2010, CMS collaborators steadily uploaded 5,000 documents into DocDB. In just one year since then, 25 users – representing major CMS groups – used the bulk uploader to import more than 120,000 documents.

Many of these documents contain details of the CMS design, construction, performance and analysis, and are vital to the future success of the CMS research programme.

4. Achieving goal 2: "Create a Coherent Web Interface to all CMS Information"

The CMS information architecture has to address the *organization, labelling, search* and *navigation* of documents and web content. Storing documents in document database systems helps address the information organisation and labelling but does not automatically give coherent search and navigation of the information. Nor does it address web pages on the estimated 1000 CMS "official" websites, many of which overlap, are out-of-date, or are wrong. To compound these issues, there is a wide variety of operating systems, technologies and navigation schemes in use.

4.1. Adoption of the Drupal Web Content Management System

These issues led us to the realisation that we needed a single, coherent web portal to all CMS information, built using a modern web content management system[3]. This coincided with a decision by CERN to follow the same strategy for the CERN websites. Following a wide consultation with departments and experiments, CERN IT department then decided to provide a new web content management service based on the Drupal[4] system, which is already used by many large organisations including the BBC, CNN, The Economist, MIT, the UN, and the White House.

Drupal is an extendable, open-source product built upon a "LAMP stack" of Linux, Apache, MySQL, and php/perl, with the following key features:

- Web content (web pages, images, etc.) is typically stored on a centrally managed file system. The associated metadata (content type, author, etc.) is stored in a MySQL relational database;
- Templated views of content (e.g. news items) and navigational elements (e.g. menus) are built from the metadata, thereby guaranteeing site-wide coherence;
- Style (layout, colours, etc.) is kept apart from content, resulting in a uniform look and feel;
- A web interface enables administrators to manage the site and users to manage their content;
- Integration with CERN single-sign-on (SSO), e-groups and Drupal permissions management ensures access is secure and that only authorised users can view or update each item;

Based on this Drupal service, we established a new collaboration website (http://cern.ch/cms) to serve all our users: collaboration members, non-CMS scientists, the public, the media, educators and students. In addition to the Drupal "LAMP" software service from CERN/IT, we added a layer of CMS-specific software: HTML templates and CSS files for customising the style and layout; new data "content types" and views; and php and JavaScript code to extend functionality, for example to manage CMS workflows, to interface to Indico, or to perform cross-system searches.

4.2. Strategic choices for the new CMS web portal

Based on numerous discussions and reviews of other leading scientific web portals, we made a number of strategic choices, as described below.

The main web content is hosted within Drupal. External web pages from other sites may also be served *via* the Drupal portal. We use HTML iframes to respect the security model of the embedded pages. A disadvantage of this is that, due to same origin security policy[5], the iframe elements cannot know the actual size of their contents, which are from a different origin. This results in undesirable scroll-bars. We circumvent this using the easyXDM[6] library to transmit the height of the embedded page to our page. This happens on the client-side as the browser frames are communicating directly (in contrast, a server-side solution involving the transfer of users' access rights would be more complex and could impose unwanted restrictions on the content accessible).

Document files are stored in document management systems, as described above, and made available through Drupal interfaces as appropriate.

The internal collaboration web content reflects the "organisational units" (groups, projects, etc.) in CMS, as shown in Figure 4. The entry page for each organisational unit conforms to a standard template (Figure 5) built from up to about 100 metadata items and links, for example: group name, coordinators, contacts, calendars, meetings, minutes, plans, documents, wikis, mailing lists, and news. Authorised users manage this information using the Drupal management interface to the underlying MySQL database. The main content of each group's entry page is under the control of the group members and may either be stored in Drupal or embedded from an external site such as a Twiki.

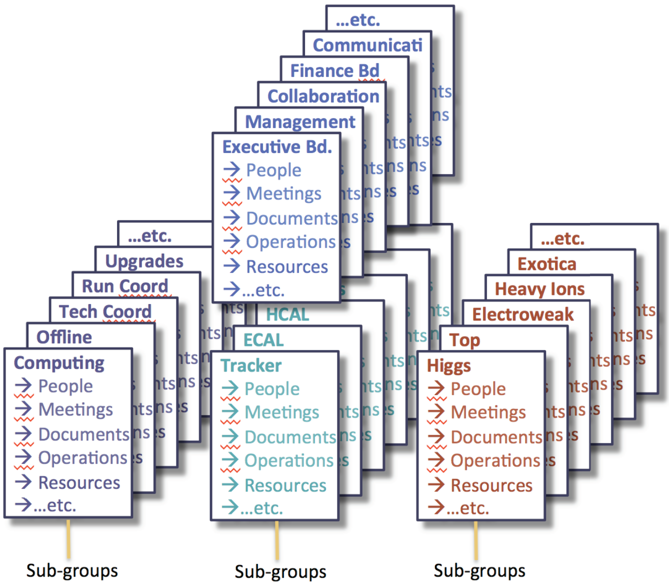
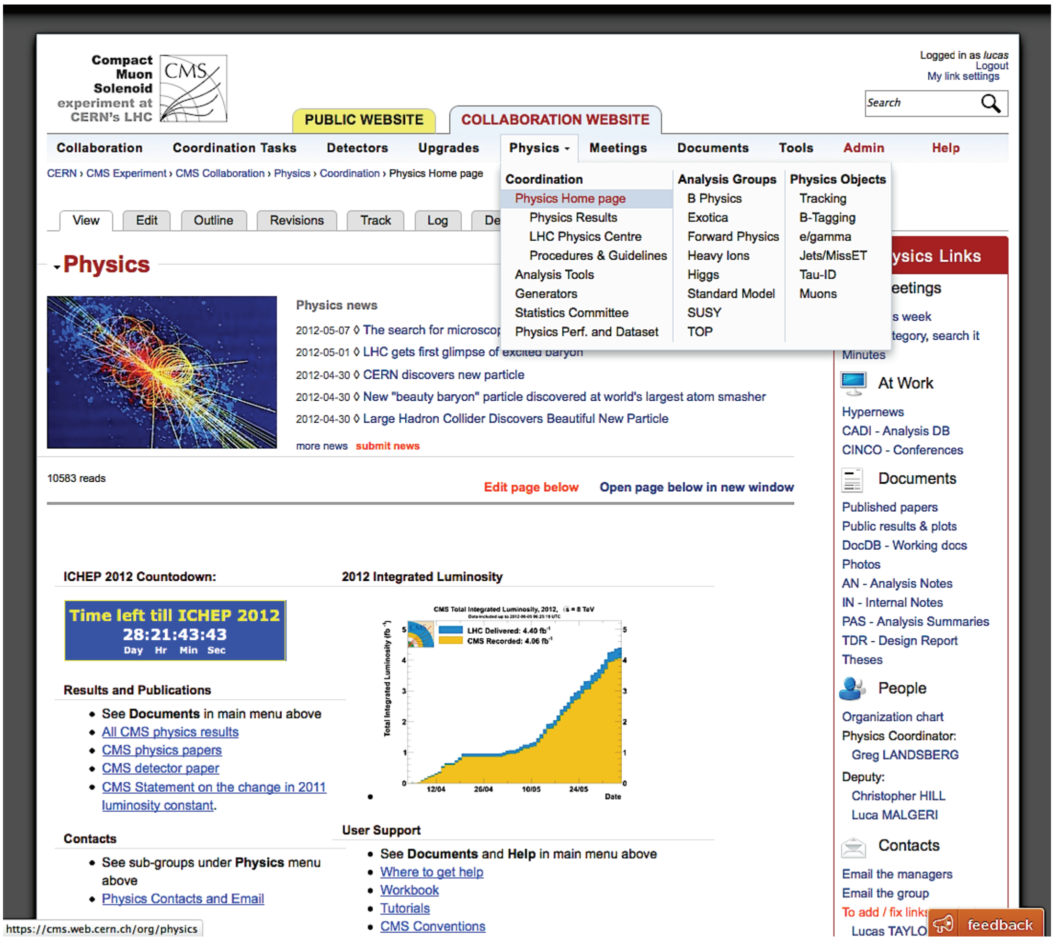


Figure 4. Schematic overview of the CMS collaboration structure.



Physics

Physics news

2012-05-07 ◊ The search for microscop
2012-05-01 ◊ LHC gets first glimpse of excited baryon
2012-04-30 ◊ CERN discovers new particle
2012-04-30 ◊ New "beauty baryon" particle discovered at world's largest atom smasher
2012-04-30 ◊ Large Hadron Collider Discovers Beautiful New Particle

10583 reads

ICHEP 2012 Countdown:

2012 Integrated Luminosity

Results and Publications

- See **Documents** in main menu above
- All CMS physics results
- CMS physics papers
- CMS detector paper
- CMS Statement on the change in 2011 luminosity constant.

Contacts

- See sub-groups under **Physics** menu above
- Physics Contacts and Email

User Support

- See **Documents** and **Help** in main menu above
- Where to get help
- Workbook
- Tutorials
- CMS Conventions

ysics Links

eetings

s week
tegory, search it
Minutes

At Work

Hypernews
CADI - Analysis DB
CINCO - Conferences

Documents

Published papers
Public results & plots
DocDB - Working docs
Photos
AN - Analysis Notes
IN - Internal Notes
PAS - Analysis Summaries
TDR - Design Report
Theses

People

Organization chart
Physics Coordinator:
Greg LANDSBERG
Deputy:
Christopher HILL
Luca MALGERI

Contacts

Email the managers
Email the group
To add / fix links
Lucas TAYLO

Figure 5. Screenshot of a typical organisational unit, showing site-wide search, top menus, group-specific news and links (right sidebar), and main page content.

Coherent site-wide navigation is built using metadata associated to the content in Drupal. Even if the content itself is elsewhere (in CDS, DocDB, etc.), the Drupal site still provides coherent *navigation* to it. A site-wide top menu bar gives access to the organisational unit structure of CMS, as well as frequently used websites and services. Different navigational views of the same metadata are easy to create, for example, a document repository index arranged by type and organisational unit.

Powerful cross-system search is provided by a CMS service built on Drupal. A single search bar in the template web page enables users to enter a free text search query. We then send this query to a dozen target systems (AIS, CDS, CERN Search, DocDB, Drupal, EDMS, Google, Indico, INSPIRE, IPPOG, phonebook, Twiki) and present the search results to the user in a coherent tabbed web page containing an iframe for each target system. A benefit of this scheme is that authentication is handled directly by the browser and the target system, without Drupal involvement, so that users can only see results that they are authorised to see.

The public part of the website serves the main non-CMS audiences of other scientists, the public, the media, educators and students, as shown in Figure 6. It is built using three main types of content: up-to-date news items (Figure 7), introductory web pages (including interactive views of the collaboration and physics results), and multimedia resources such as fact sheets, brochures, images, video clips, and event data for outreach and education. We added workflow management in Drupal for the writing, editing and approval of news items, and their distribution via the website, Email, RSS and social media platforms such as Google+, Facebook, Twitter. For multimedia items, Drupal acts as a thin portal to content stored in document management systems and well established and professional quality external sites such as YouTube.

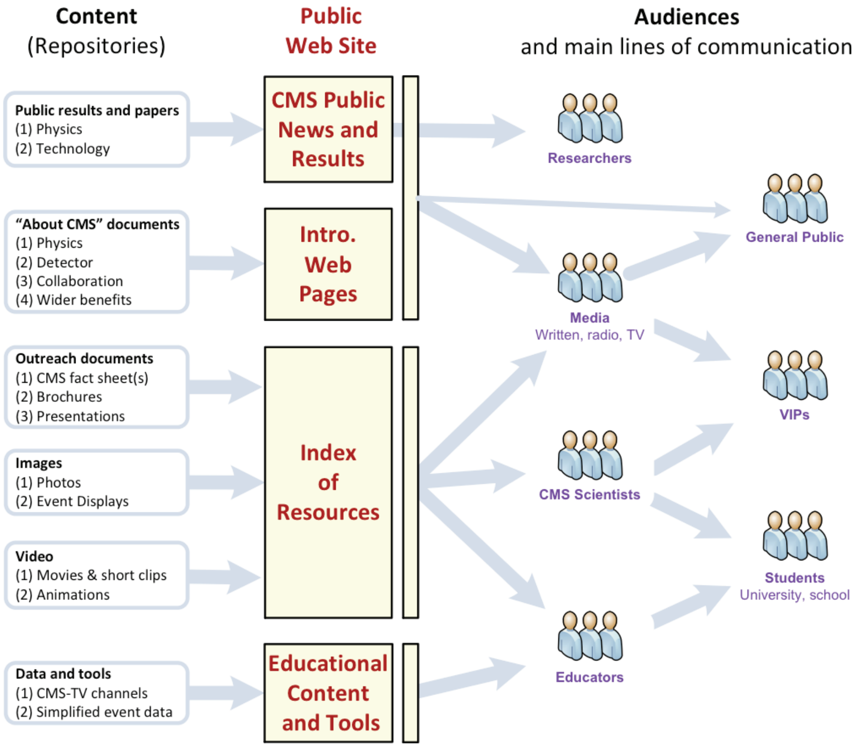
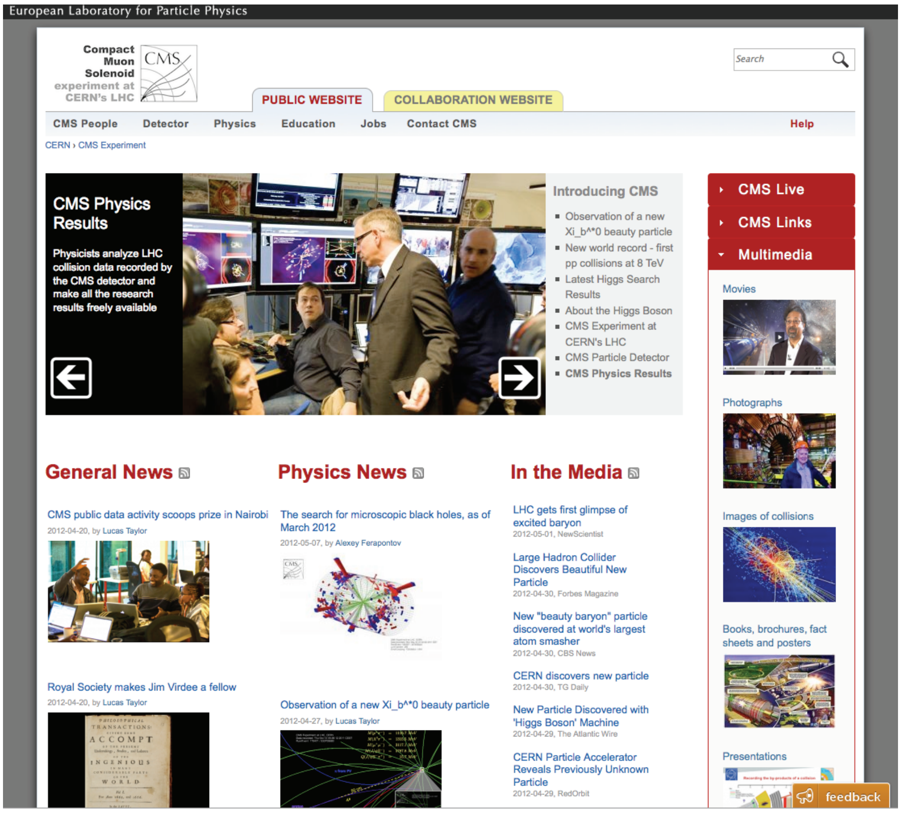


Figure 6. Schematic overview of the CMS public website content and audience.



CMS Physics Results

Physicists analyze LHC collision data recorded by the CMS detector and make all the research results freely available

Introducing CMS

- Observation of a new Xi_b^*0 beauty particle
- New world record - first pp collisions at 8 TeV
- Latest Higgs Search Results
- About the Higgs Boson
- CMS Experiment at CERN's LHC
- CMS Particle Detector
- CMS Physics Results

General News

CMS public data activity scoops prize in Nairobi
2012-04-20, by Lucas Taylor

Royal Society makes Jim Virdee a fellow
2012-04-20, by Lucas Taylor

Physics News

The search for microscopic black holes, as of March 2012
2012-05-07, by Alexey Ferapontov

Observation of a new Xi_b^*0 beauty particle
2012-04-27, by Lucas Taylor

In the Media

LHC gets first glimpse of excited baryon
2012-05-01, NewScientist

Large Hadron Collider Discovers Beautiful New Particle
2012-04-30, Forbes Magazine

New "beauty baryon" particle discovered at world's largest atom smasher
2012-04-30, CBS News

CERN discovers new particle
2012-04-30, TG Daily

New Particle Discovered with 'Higgs Boson' Machine
2012-04-29, The Atlantic Wire

CERN Particle Accelerator Reveals Previously Unknown Particle
2012-04-29, RedOrbit

Figure 7. Screenshot of the public part of the CMS website showing the strong focus on latest news, background information, and multimedia resources.

4.3. *Migration to the new CMS web portal*

In order to ensure a smooth transition to the new web portal, we considered the constraints of the main stakeholders carefully and addressed them, as follows.

CMS collaborators and groups accept the need to improve the organisation of their documents but can rarely commit much time to do it. Therefore the new document systems (DocDB and the bulk uploader) and the web portal were designed to be easy and self-explanatory, such that users naturally choose to migrate without any external pressure or the need for training. Since we can only encourage but not enforce a standard way of doing things, the information systems were designed to satisfy a range of needs within a common, coherent framework. An example of this is the facility for embedding legacy web pages into the standard framework and navigation scheme of the new site.

CERN IT services are used wherever possible to minimise the CMS effort required and to leverage CERN expertise. Since the CERN service providers must satisfy a diverse community beyond CMS, we do not expect them to significantly customise their services to meet specific CMS needs.

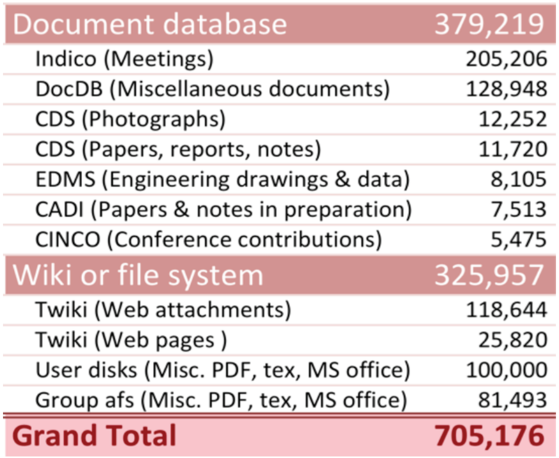
The CMS Communications Group has very limited personnel. Its top priority was to create a coherent and robust web portal framework that included all CMS groups and users in a coherent fashion. The huge job of cleaning up legacy content was, of necessity, assigned a lower priority.

Before rolling out the new website, we systematically contacted key people in all the main 60 CMS groups and projects to get feedback and help filling in missing metadata (e.g. links for their own projects). Only once we felt that the new website's content and navigation was more much complete and correct than the existing CMS websites did we encourage all CMS users to try the new services. In parallel, all the pre-existing systems were kept running without any changes. We strongly encouraged user feedback, through a prominent widget on all web pages, and we monitored the access patterns for old and new websites using Google Analytics[7].

5. Current Status and Future Plans

5.1. Status of CMS document management

Table 2 shows the total number of CMS documents and their breakdown by location. More than half of the total (380,000 documents) are now in a document management system, representing an increase of about 100,000 documents since the start of our work. Several hundred thousand documents remain to be processed and a significant fraction of these will need to be uploaded into DocDB.



Document database	379,219
Indico (Meetings)	205,206
DocDB (Miscellaneous documents)	128,948
CDS (Photographs)	12,252
CDS (Papers, reports, notes)	11,720
EDMS (Engineering drawings & data)	8,105
CADI (Papers & notes in preparation)	7,513
CINCO (Conference contributions)	5,475
Wiki or file system	325,957
Twiki (Web attachments)	118,644
Twiki (Web pages)	25,820
User disks (Misc. PDF, tex, MS office)	100,000
Group afs (Misc. PDF, tex, MS office)	81,493
Grand Total	**705,176**

Table 2. Numbers of CMS documents in document systems (May 2012), the Twiki and on user and group file systems.

5.2. Status of the new CMS web portal

The new CMS web portal became a full production service in the autumn of 2011. CMS and public users quickly discovered the new site before it was even announced and started using it, as shown in Figure 8. By the time the new site was officially announced in Jan 2012, it was already receiving 84% of the combined traffic to the old and new sites. This fraction is now 96% meaning the old sites can soon be deprecated.

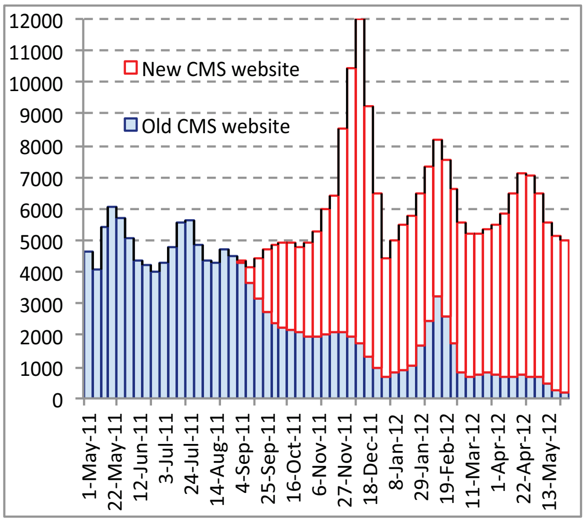


Figure 8. Numbers of visits per week, between May 2011 and May 2012, to the old CMS public website (lower, blue) and the new (upper, red) CMS web portal.

5.3. Future work

Although the new web portal has been very well received so far, we know that it can always be improved. Therefore, in February 2012 we conducted a CMS user feedback survey, to which we received 306 responses, or approximately 10% of CMS collaborators. The full report of the survey and the associated action items are described in Reference 8. In summary, the main tasks for the near future are to:

1. Improve the interfaces and tools for finding and working with CMS documents;
2. Import the estimated 180,000 documents from disks and websites into DocDB, and then archive and retire websites that are no longer required;
3. Provide interfaces and feeds for (Indico) meetings, (Google) calendars, conferences and news;
4. Add live CMS operations displays (LHC and CMS pages 1, luminosity, event display; and
5. Develop a strategy for improving the structure and quality of 144,000 Twiki pages and files.

5.4. Some subjective advice

We close with some subjective advice for others who may be facing a similar challenge:

- Ask yourself honestly: “How much information do we really have?” “Can I afford to lose it?” “Is it in a secure system?” “Can I find what I want, when I need it?”
- Everybody agrees that improving information systems is important but, to encourage people to help, you need systems that are trivial to use and that automate repetitive tasks.
- Adopt a web Content Management System for any reasonably sized website. The extra initial effort will soon pay off. Keep your content data types simple and, to retain generality, don’t define too many similar ones.
- Drupal can probably meet the content management needs of most HEP collaboration websites. If, however, you need to customize it extensively you should also consider other web tools.
- To make an effective transition, ensure the new website is easier to use and more complete than the old one(s). Focus on coherent design and navigation. Embed legacy content if need be, at least initially. Keep the old sites running in parallel until the transition is complete, and then retire them. Use Google Analytics to monitor access patterns.
- The skills and costs are very modest compared to the significant benefits. To paraphrase Derek Bok, “if you think managing information is expensive, try not managing it!”

6. References

[1] Private communication, T. Camporesi, M. Chamizo-Llatas (former/current Run Coordinators)
[2] https://cms-docdb.cern.ch/cgi-bin/PublicDocDB/DocumentDatabase
[3] http://en.wikipedia.org/wiki/Content_management_system
[4] http://drupal.org/
[5] http://en.wikipedia.org/wiki/Same_origin_policy
[6] http://easyxdm.net/
[7] http://www.google.com/analytics/
[8] https://cms-docdb.cern.ch/cgi-bin/PublicDocDB/ShowDocument?docid=5960