



On the usage of Python in the CERN Document Server's digital library and conference management tools



Why this presentation ?

- All CDS (CERN Document Server) applications are using Python for
 - Management of events/conferences: Indico
 - Management of documents: Invenio
- Europython is using CDS Indico to help managing this conference
- Europython at CERN



Content

- CDS Indico & Invenio
 - Overview of the software features
- Technologies and Licensing at CDS
- Python at CDS
 - Why was Python selected ?
 - How good/bad is our experience ?
- Conclusion



Managing Documents with **invenio**



What is Invenio ?

- CDS Invenio software is a **document repository application** that enables to run an electronic preprint server, a digital library catalogue or a document archive on the web
- At CERN, we use it for:
 - High Energy Physics e-archive
 - Institutional scientific repository with documents, photos, videos and more
 - About 1 million records; 500 collections; 200,000 users/year
 - designed to cope with new dissemination channels of scientific results of LHC (Open Access)
- tries to combine the best of traditional Library world and modern information retrieval technologies
- uses existing standards, e.g. the US Library of Congress standard to describe documents, Unicode, OAI, etc.



Some features (I)

■ Navigable collection tree

- Documents organised in collections
- Regular and virtual collection trees
- Customizable portalboxes for each collection

■ Powerful search engine

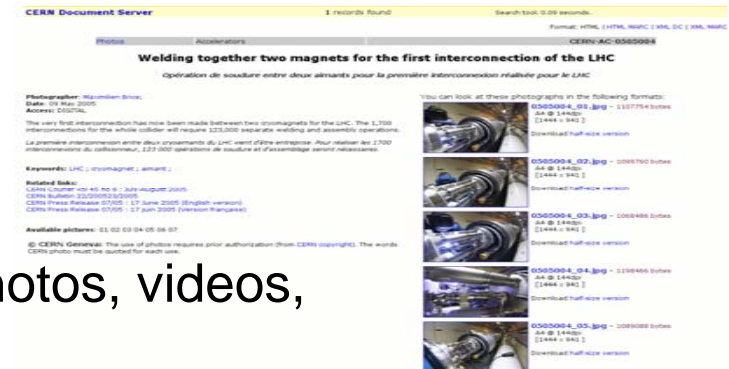
- Specially designed indexes to provide Google-like search speeds for repositories of up to 1,500,000 records
- Customizable simple and advanced search interfaces
- Combined metadata, fulltext and citation search in one go
- Results clustering by collection
- Interface in 16 languages





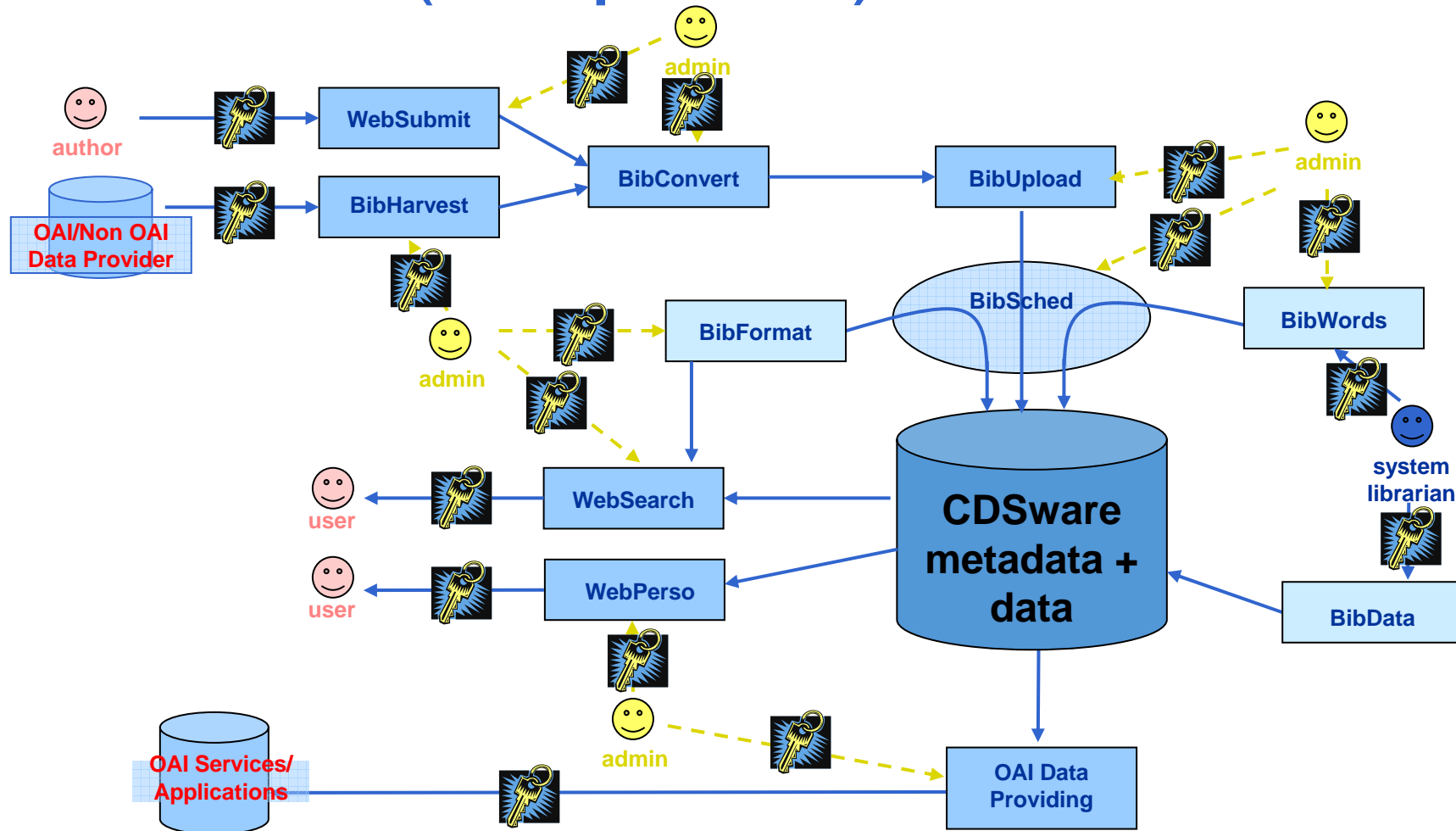
Some features ? (II)

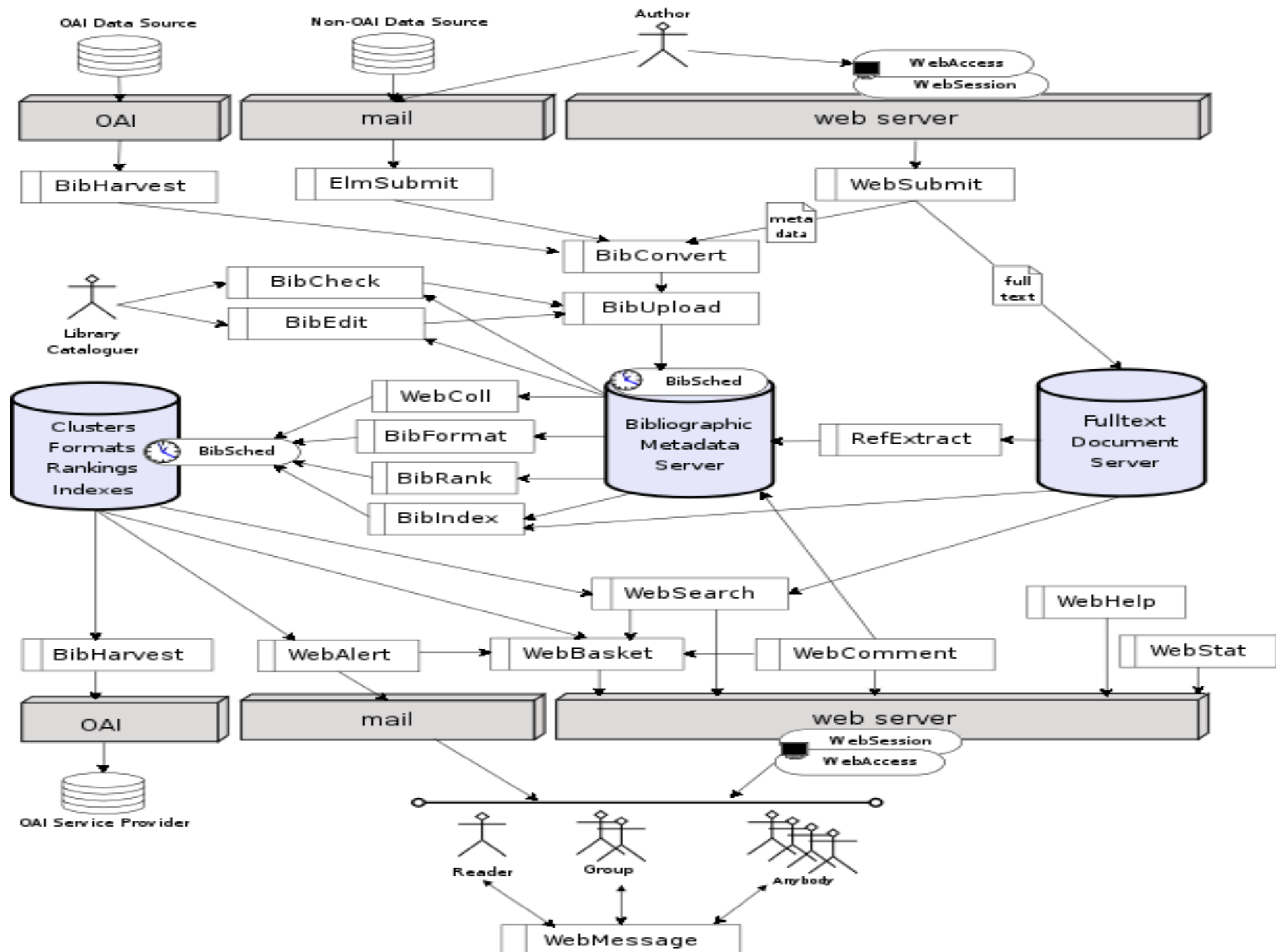
- **Flexible metadata**
 - Standard metadata format (MARC)
 - Handling articles, books, theses, photos, videos, museum objects and more
 - Customizable display and linking rules
- **Collaborative tools**
 - user-defined document baskets & automated email notification alerts
 - basket-sharing within user groups
 - user comments and reviews of documents





Invenio (simplified) view







CERN Document Server

Search:

NA48 any field

[Search Tips](#) :: [Advanced Search](#) :: [Try your search on...](#)

Search collections:

*** any collection ***

Sort by:

- latest first - desc. - or rank by -

Display results:

10 results split by collection

Output format:

HTML brief

- HTML address label
- HTML brief
- HTML detailed
- HTML MARC
- HTML photo captions only
- HTML portfolio
- XML Dublin Core
- XML MARC

Results overview: Found 365 records in 0.31 seconds.

[Articles & Preprints](#), 301 records found

[Books & Proceedings](#), 8 records found

[Presentations & Talks](#), 3 records found

[Multimedia & Outreach](#), 46 records found

[Archives](#), 7 records found

Articles & Preprints

301 records found 1 - 10 jump to record:

- 1. **Rare Kaon Decays $K \rightarrow \pi n n$ and $KL \rightarrow \pi l l$** / [Mescia, F](#)
Over the next years, the Flavour Physics community will be looking for inconsistencies of the Standard Model (SM) by exploiting new and precise measurements. [...] [hep-ph/0512142](#); 12 Dec 2005. - mult. p [Fulltext](#)
[Detailed record](#) - [Similar records](#)
- 2. **Size of isospin breaking in charged $K/\text{sub } 14/$ decay** / [Nehme, A](#)
We evaluate the size of isospin breaking corrections to form factors f and g of the $K/\text{sub } 14/$ decay process $K/\text{sup } +/ \text{ to } \pi/\text{sup } +/ \pi/\text{sup } -/ \text{sup } +/ \nu l$ which is actually measured by the extended NA48 setup at CERN. [...] 2005 - Published in: [Eur. Phys. J., C 40 \(2005\) 367-82](#)
[Detailed record](#) - [Similar records](#)
- 3. **Direct CP Violation in Charged Kaon Decays by the NA48/2 Experiment at CERN** / [Collazuol, G](#)
The NA48/2 experiment at CERN is performing high precision studies of charged kaon decays, using an upgraded NA48 setup and a novel design for simultaneous unseparated $K/\text{sup } +/ \text{ or } -/$ beams. [...] 2005 - Published in: [Nucl. Phys. B, Proc. Suppl. 142 \(2005\) 293-8](#)
[Detailed record](#) - [Similar records](#)
- 4. **Observation of a cusp-like structure in the pizero-pizero invariant mass distribution from $K^+ \Rightarrow \pi^+ \text{ pizero pizero}$ decay and**



CERN Accelerator R&D Projects

Search 40 records for:

[Search Tips](#) :: [Advanced Search](#)

EU co-funded projects:



CARE

- **Locally available documentation:**
 - [CARE Introduction](#)
 - [CARE Workshops](#)
 - [CERN contributions and reports](#)
 - [CERN Resources for CARE](#)
- **External documentation:**
 - [CARE website](#)
 - [CARE Agenda](#)
 - [CARE Notes and Reports](#)
 - [CARE Annual Report](#)
 - [CARE Meetings and Minutes](#)

DIRAC

- **Locally available documentation:**
 - [DIRAC Introduction](#)
 - [DIRAC Workshops](#)
 - [CERN contributions and reports](#)
 - [CERN Resources for DIRAC](#)
- **External documentation:**
 - [DIRAC website](#)
 - [FAIR website](#)



EURISOL

- **Locally available documentation:**
 - [EURISOL Introduction](#)
 - [EURISOL Workshops](#)
 - [CERN contributions and reports](#)
 - [CERN Resources for EURISOL](#)
- **External documentation:**
 - [EURISOL website](#)



CERN PhotoLab

Search:

[Search Tips](#) :: [Advanced Search](#) :: [Try your search on...](#)

Search collections:

Sort by:

Display results:

Output format:

CERN PhotoLab

853 records found 1 - 50 jump to record:

Search took 0.14 seconds.





Managing Events with





CERN Document Server software

JY. Le Meur; T. Baron

T. Simko; D. Bourillot

Europython – 4th July 2006



Europython 2006

3-5 July 2006

CERN, Geneva

[Home](#)

- ▶ General Information
- ▶ **Europython Home**
- ▶ Programme
- ▶ Timetable
- ▶ Speaker list

Call for Abstracts

- ▶ View my abstracts

✉ [support](#)

REGISTRATION ↑

REGISTRATION
NETWORK
TRAVEL



Number of events:

Year	Number of events
1995	19
1996	3
1997	5
1998	25
1999	138
2000	354
2001	676
2002	1552
2003	2480
2004	3621
2005	5714
2006	4456
2007	69
2008	25
2009	2

Total: **19139**





Project History

- **Indico (Integrated Digital Conference)**
 - European project: 2002-2004
 - Partners:
 - Italy: SISSA, University of Udine
 - Holland: TNO TPD, University of Amsterdam
 - CERN
 - In production at CERN since 2004 (first time use: CHEP'2004)
 - Currently hosts >100 conferences
 - Usage is growing fast
 - <http://indico.cern.ch>

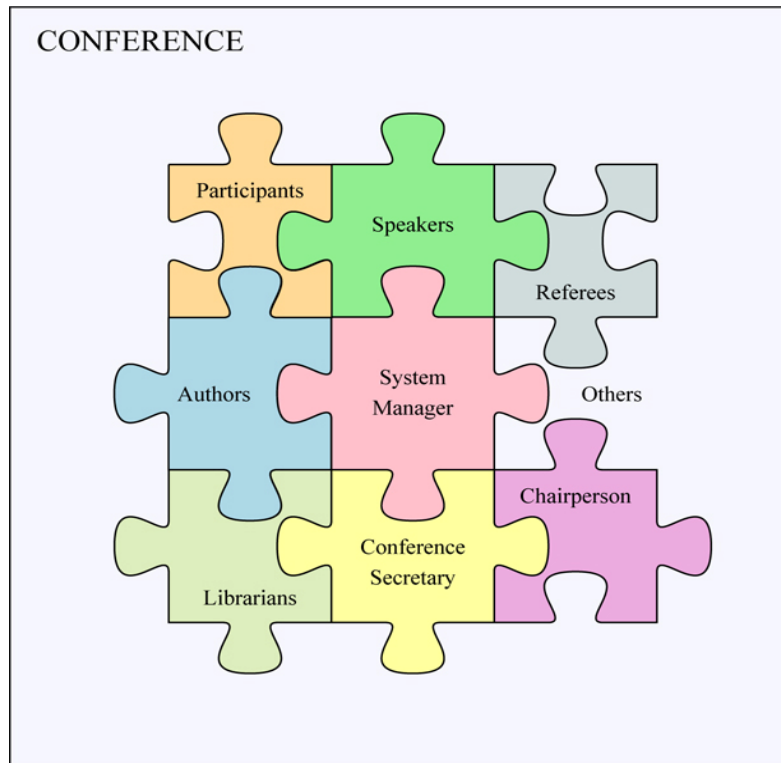


Conference Management

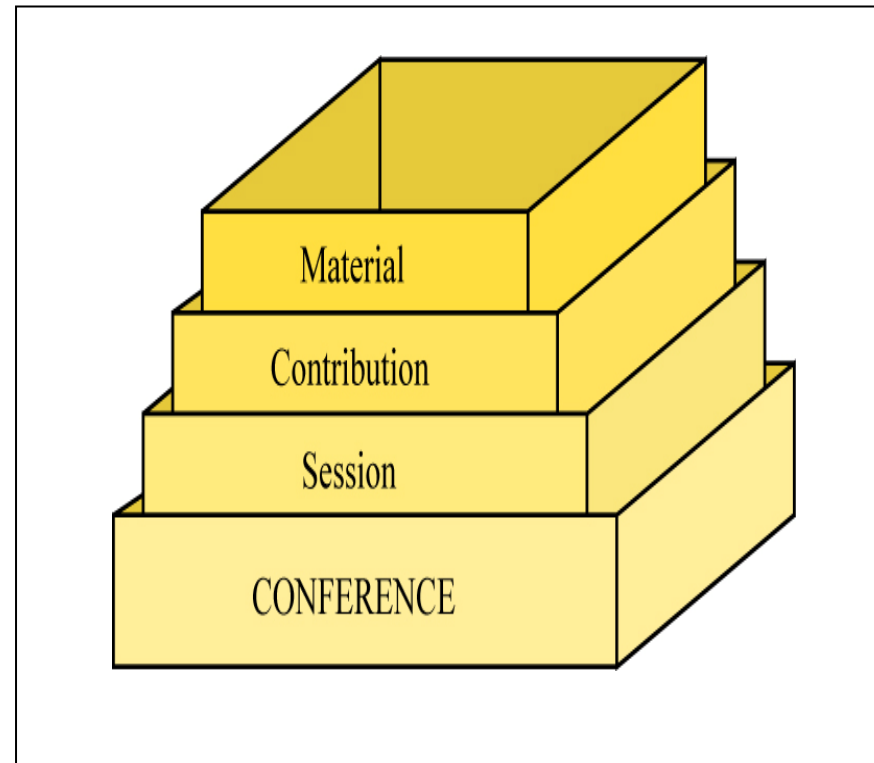




- A complex event...



human



logical



- ...with a lot of processes

- > Web Site Management
 - > Conference Program Definition
 - > Attendant Registration
 - > Call for Abstract
 - > Abstract Selection
 - > Material Submission
 - > Timetable Setup
 - > Electronic Proceedings
 - > Long term archive
-



CERN Document Server software

JY. Le Meur; T. Baron

T. Simko; D. Bourillot

Europython – 4th July 2006



login



BARON, Thomas - logout



Management Area

Integrated Digital Conference

Home > Conferences > International Workshop on African Research & Education Networking



Conference International Workshop on African Research & Education Networking

- Main
- Programme
- Contributions
- Timetable
- Book of Abstracts
- Call for Abstracts
- Abstracts
- Access Control
- Archiving
- Registration Form
- Registrants
- Display
- Listings
- Tools

Title International Workshop on African Research & Education Networking

Description The workshop is jointly organized by CERN, ITU and the United Nations University (UNU) in close cooperation with DANTE, IDRC, IEEAF, Internet2, ISOC, RENATER, TERENA and UNESCO. The Norwegian Agency for Development (NORAD) is funding a feasibility study for the African University Network (AFUNET) as a practical response to the World Summit on Information Society (WSIS) Plan of Action. AFUNET's main objective is to enhance the capabilities of African academic and scientific institutions to take advantage of the opportunities associated with the emergence of the global information society.

Place CERN

Geneva, Switzerland

Room: Council Chamber

Start date Sunday 25 September 2005 08:00

End date Tuesday 27 September 2005 18:00

Additional info

Support email olivier.martin@cern.ch

Default style --not set--

Visibility up 999 levels

modify

Chairpersons Dr. HOFFMANN, Hans Falk

remove

Management features

Enabled feature
 Disabled feature

Call for abstracts
 Registration Form

Status

close



- Workshop Website
- List of registrants
- Author index
- Overview**
- Contribution List
- Speakers Index
- Scientific Programme
- Timetable
- Book of abstracts [PDF]

support



Meeting Management





CERN Document Server software

JY. Le Meur; T. Baron

T. Simko; D. Bourillot

Europython – 4th July 2006

- Less actors, processes, complexity
- Same core, simplified interfaces



category | view: Indico style | manage ©BARON, Thomas - [logout](#)

INDICO Integrated Digital Conference **Management Area** ©BARON, Thomas - [logout](#)

[Home](#) > [Departments](#) > [IT](#) > [Groups](#) > [UDS](#) > [Group Meeting](#) > [IT-UDS Group Meeting](#)

Meeting **IT-UDS Group Meeting**

[Main](#) | [Timetable](#) | [Access Control](#) | [Listings](#) | [Tools](#)

Title IT-UDS Group Meeting
Description
Place CERN

Room: 513-1-024

Start date Thursday 25 August 2005 10:00
End date Thursday 25 August 2005 10:30

Additional info
Support email --not set-- [modify](#)

Chairpersons Text
Visibility up 999 levels
Default style CDS Agenda style

Chairpersons **SMITH, Tim** [remove](#)

Time Table:

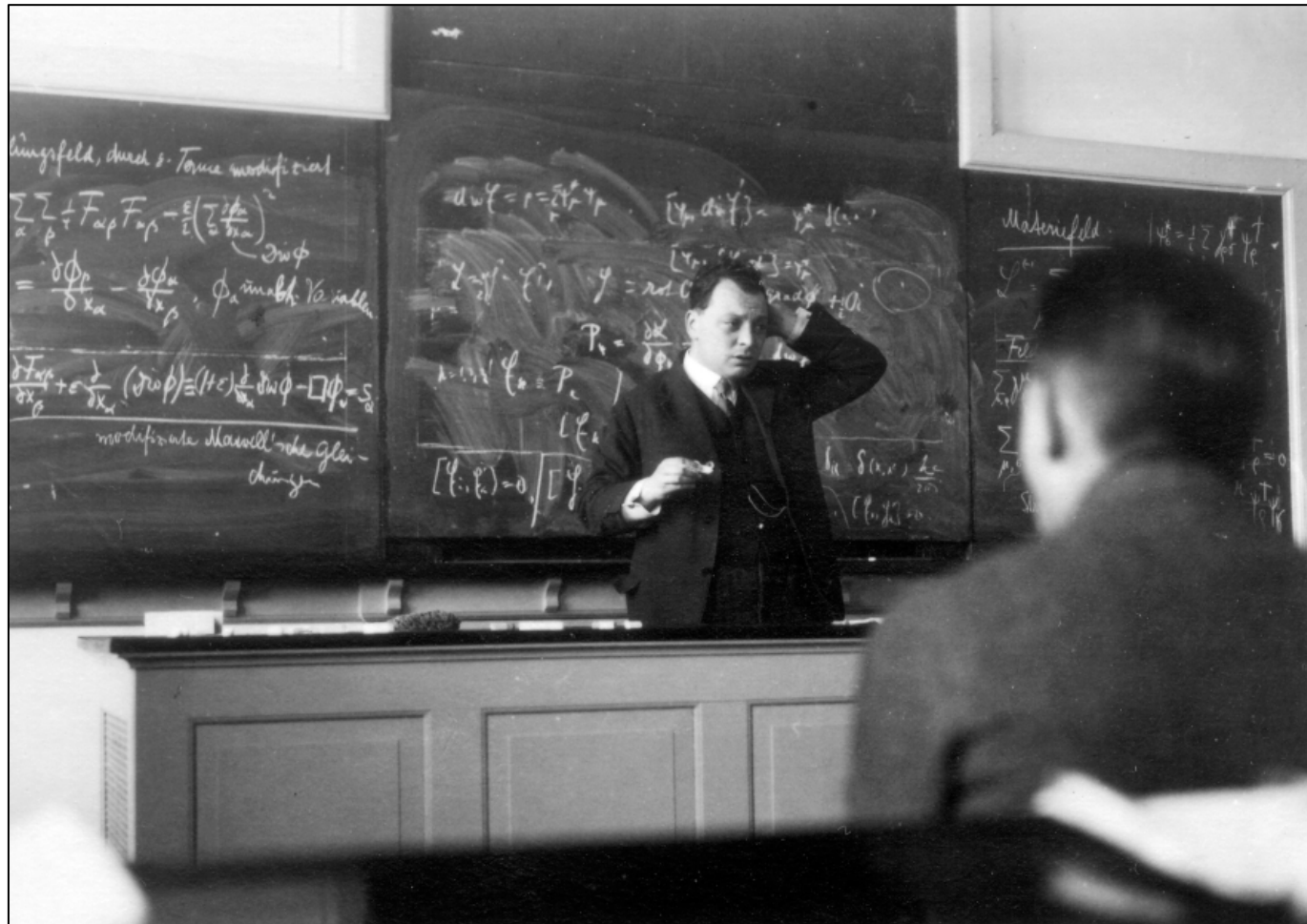
Thursday 25 Au

10:00	Introduction
10:10	Video Con
10:20	Discussion
10:25	Using Illustr
10:35	Discussion
10:40	Supporting
10:50	Discussion
10:55	IT BookSho
	transparenci
11:05	Discussion

CERN | Powered by [IndiCo 0.8.2](#) |



Lecture Management





CERN Document Server software

JY. Le Meur; T. Baron

T. Simko; D. Bourillot

Europython – 4th July 2006



category ▶▶ | view: Lecture | manage ©BARON, Thomas - logout

Management Area ©BARON, Thomas - logout

Home > Departments > IT > Groups > UDS > Presentations > CDS Search Technology

DISPLAY Lecture CDS Search Technology

Main | Access Control | Tools

Title CDS Search Technology

Description In a first part, Jean-Yves Le Meur introduces CERN Document Server Search module, its history, scope and software architecture. In a second part, Tibor Simko details the search technology itself, its index design and some performance issues.

Place CERN

Room: 513-1-024

Start date Friday 06 February 2004 09:30

End date Friday 06 February 2004 10:30

Additional info

Support email --not set--

Chairpersons Text

Default style Lecture

Visibility up 999 levels



Planning/Archiving

- One server – Many events of various sizes
- Hierarchical organisation: tree of categories to classify the events
- Search engine provided by CDS Invenio through an OAI harvesting

The screenshot shows the INDICO web interface. At the top left is the INDICO logo and 'Integrated Digital Conference'. At the top right is the user 'BARON, Thomas' with a 'logout' link. The main content area is titled 'Category Map for Home' and displays a hierarchical tree of categories:

- **Committee Meetings**
 - Finance Review Committee of CAST
 - Finance Review Committee of COMPASS
 - Joint CERN-Pakistan Committee
 - CERN-Russia Joint Working Group
 - CERN-US Committee
- **Conferences**
- **Departments**
 - ETT
 - ETT-DH Meetings
 - Audio-Video
 - CDS Section Meetings
 - Section Leaders Minutes
 - miscellaneous
 - IT
 - test
 - Groups
 - UDS
 - AVC section
 - CDS Section
 - CERN Serco Management Review Meetings
 - Group Meeting
 - Presentations
 - Section Leaders Meeting
- **Experiments**
 - CMS Meetings
 - CMSCC
 - SPIE International Congress On Optics And Optoelectronics, Photonics Applications In Industry And Research (PA-IV)- (Photonics Applications in Industry and Research)
- **TEST Category**

On the right side of the interface, there is a 'Tools' menu with the following items:

- ▶ Display
- ▶ Overview
- ▶ Calendar
- ◀ Map
- ▶ Modify
- ▶ Admin
- ▶ Statistics



Planning/Archiving

INDICO Integrated Digital Conference

Overview for Home

Period: Detail:

Key:

<< < Monday 26 September 2005 >>

08:00	- Pixel part of ID week (Kevin Einsweiler) (CERN)
09:00	- ATLAS Software & Computing Workshop (Dario Barberis & David Quarrie) (CERN)
09:00	- IMMW 14 (CERN - Geneva - Switzerland)
09:00	- ID week (Leonardo ROSSI) (CERN)
13:00	- COD-4 Meeting (P. Strange/H.Cordier) (Cosener's House, Abingdon)
14:00	- Software week (Nick Brook) (CERN) (2-r-30)
14:00	- Operations meeting (Nick Thackray) (CERN)
14:00	- Operations meeting (Nick Thackray) (CERN)
14:00	- JRA3 status meeting (EDLUND, Ake) (Phone meeting)
15:00	- Atlas Monte Carlo meeting (Hinchliffe, I.; Kersevan, B) (CERN) (40-R-B10)
16:00	- Data Quality Monitoring (BARTALINI, Paolo) (CERN) (40-4-C01)

overview

INDICO Integrated Digital Conference

Calendar

Options

Display next months over columns starting

Color Legend

Committees Conferences Departments Experiments Projects Multiple

September 2005							October 2005						
Mo	Tu	We	Th	Fr	Sa	Su	Mo	Tu	We	Th	Fr	Sa	Su
			1	2	3	4					1	2	
5	6	7	8	9	10	11	3	4	5	6	7	8	9
12	13	14	15	16	17	18	10	11	12	13	14	15	16
19	20	21	22	23	24	25	17	18	19	20	21	22	23
26	27	28	29	30			24	25	26	27	28	29	30
							31						

calendar



Summary

- Supports full event lifecycle:
 - Preparation of the event
 - Live usage for accessing agenda & stored material
 - Long-term archival of the events information and related files
- Typical Use Cases
 - [Conferences](#)
 - [Workshops](#)
 - [Meetings](#)
 - [Seminars](#)



Technologies and Licensing at CDS



The Indico Technology

- Main programming language: *Python*
- Runs on Apache using the Python module *mod_python*
- Persistence based in **ZODB** (Zope Object Database)
 - Transparency: no need for explicit read/writes of the objects
 - Fits very well with Indico complex object model
 - Proven performance and scalability
- Timetable generation: libXML, libXSLt + python bindings
- Portable technologies: runs on Windows, linux
- Export gateways:
 - iCalendar ; XML ; PDF outputs
 - **OAI** (Open Archive Initiatives) for ensuring integration with other services
 - Standard protocol for information exchange between digital libraries
 - Allows to expose conference data
 - Allows other systems to fetch conference data and build services over it
 - Simple mechanism → XML over HTTP



The Invenio Technology

- Main programming language: *Python*
- Runs on Apache using the Python module *mod_python*
- Uses MySQL RDBMS
 - Take advantage of fully featured query language
- Invenio home made Indexes
- Internal representation with XML-MARC
- Export gateways:
 - Multiple output formats: HTML, XML, MARC, OAI, DC, etc.
- Some modules:
 - Still in PHP (slowly moved to Python)
 - Some in Common Lisp (BibCheck)



Licensing - conditions

- GNU GPL
- Regular public releases of software packages
- Support modes
 - Free via listboxes
 - Charged
- CDSware Development Consortium
 - Main partners: EPFL, EIF; exchanging students, code, strategy
 - World wide contributions; internationalization
 - Open to newcomers !



Licensing - installations

- Invenio:
 - *HBZ NRW (Köln Germany),*
 - *Università La Sapienza (Rome, Italy),*
 - *Aristotle University (Thessaloniki, Greece),*
 - *Université catholique de Louvain (Belgium),*
 - *UCSD (San Diego, USA),*
 - *RERO (Martigny, Switzerland),*
 - *EPFL (Lausanne, Switzerland),*
 - *Swiss Library Consortium, ETHZ (Switzerland)*
 - *Educa.ch (Swiss Education Server)*
 - *CINI Foundation (Italia)...*
- Indico:
 - *DTV (Denmark),*
 - *UIUC (Illinois, USA),*
 - *Fermilab (Chicago, USA),*
 - *EPFL (Lausanne Switzerland),*
 - *DESY (Hamburg, Germany),*
 - *U. of Mexico (Mexico),*
 - *TRIUMPH (Canada) ...*



How/Why has CDS selected Python ?



Two distinct evolutions

- **1993** - CERN Preprint Server on the web: *CERN httpd; CGI - C/Shell/Perl Programming*
- **1998** - CERN Web Library: *PHP/MySQL and C APIs to Library System*
- **2001** – CDSware starts introducing *Python/mod-python* in some components
- **2006** – CDS Invenio released with all modules in *Python*
- **1996** - CDS Agenda: *PHP and MySQL*
- **2002** - INDICO EU Project:
 - Development Process based on *Unified Software Development Process (light version)*
 - Implementation of several prototypes for validation and ensuring quality & scalability
- **2004** – CDS Indico app: *Python and ZODB*



Invenio



Indico

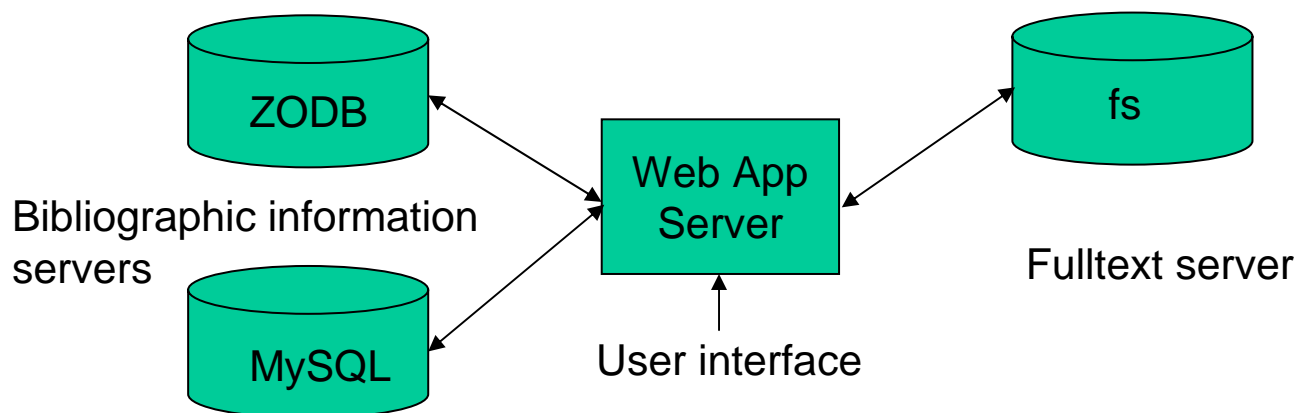


With extra applications...

- Document Format Conversion
CERN Conversion Server <http://cdsconv.cern.ch>
 - Video Analysis <http://www.eif.ch/projets/smac/>
 - Electronic Bulletins <http://bulletin.cern.ch>
 - Generation of Lists (publications, events, etc)
- Search Engine used as a *Platform*
- Considered as the heart of all the apps

Web App Server vs. DB Server

- Three-tier system architecture



- **Web App Server vs. DB Server:** which one to load?
- Native (fulltext) MySQL indexes:
 - 500,000 records ! 25+ Mrows ! 5+ sec searches
 - Google-like speed for up to 100,000 records only



Index Space Design (I)

- **Performance-driven design** assumptions:
 - low number of updates, high number of selects
 - fast searching, slow indexation
 - put load on Web App Server, free DB Server
 - cache everything cacheable
- **Search modes:**
 - search for words
 - search for phrases (exact, partial)
 - search for regular expressions
- **Index types:**
 - forward : $term1 \rightarrow [rec1, rec2, \dots]$
 - reverse : $rec1 \rightarrow [term1, term2, \dots]$



Index Space Design (II)

- Two important **speed factors** to consider:
 - speed of set intersections (Web App Server)
 - speed of set marshalling (Web App <-> DB Server)
- **Data structures** tested:
 - sorted (lists, Patricia trees)
 - unsorted (hashed sets, binary vectors)
- **fast prototyping:** (Python)
 - throw-away coding, organic-growth software
- **development model**
 - typical search time gain: 4.0 sec → 0.2 sec
 - typical indexing time loss: 7 hours → 4 days
 - *binary vectors* found the best compromise (for all types of sets)



Performance Benchmarks (2002)

- **Testing** marshalling/intersection/union/unmarshalling
- **Bytecode** interpreted language study: (Python, Java)
 - Python faster than Java (mainly due to marshalling)
- **Machine code** compiled language study: (ML, Lisp)
 - OCaml, CMU CL: 3+ times faster than Python C libs
 - CMU CL best scalable: intersecting 6M records in 0.01 sec, 30M records in 0.04 sec
- **Data structure** study:
 - OCaml, 3,000,000 records: bit vectors 0.43 sec, hashed sets 1.71 sec, lists 3.76 sec, Patricia trees do not scale well for dense sets
- **Python fast enough for production** (1M records)
 - fast C modules: Numeric (byte/bit), Marshal, Psyco



Performance Stats (2004)

- Dual Xeon(HT) 3.06 GHz, SCSI Ultra320
- 650,000+ records, 450+ collections
- **Indexing:** total index size 11 GB, indexing time 2 days
 - global words index: 3,000,000+ words
 - global words index growth rate: 2.8 words/record
 - title words index growth rate: 0.1 words/record
- **Searching:** typical search speed

<i>query</i>	<i>no. hits</i>	<i>search time</i>
ellis	1,797	0.07 sec
cern	223,843	0.07 sec
of	439,793	0.07 sec
of cern	109,635	0.10 sec
of cern the this	11,940	0.17 sec



The + of Python

- Clean aesthetical language
- Easy to learn, important for many internship students and temporary members working on the project
- Very good for rapid prototyping & organic-growth development
- Plenty of ready-to-be-used modules
- Bytecode-compiled only, speed okay for our needs



The – of Python

- No standard: danger of removing language features like *lambda* and friends (*map*, *reduce*, *filter*)
- Only basic dynamic redefinition capabilities, not like Common Lisp
- At some point, when collection size reaches a few million of documents, Python 'slowness' will be an issue...



Conclusion

- CDS Indico & Invenio are two Python applications developed at CERN running world wide
- We are satisfied with this choice, and students enjoy learning & using it
- Two reasons for a possible change:
 - Search Engine into C, OCAML or CL for performance reasons
 - Python 3000 evolution



Questions ?

<http://cdsware.cern.ch>