Invenio:
a modern digital library system

Samuele Kaplun <Samuele.Kaplun@cern.ch>
(on behalf of the Invenio Development Team)

<http://invenio-software.org/>

History I

- CERN adopted Open Access since almost **60 years**
- **Explored and invented** the tools which were needed to accomplish this (starting with *paper dissemination* in 1954)
- **Invenio** (formerly known as CDSware) answers the need of having a large scale digital library (since 2002)

From the CERN Convention:
Paris, 1st July 1953

*[…] **the results of its experimental and theoretical work shall be published or otherwise made generally available**. […]*

History II

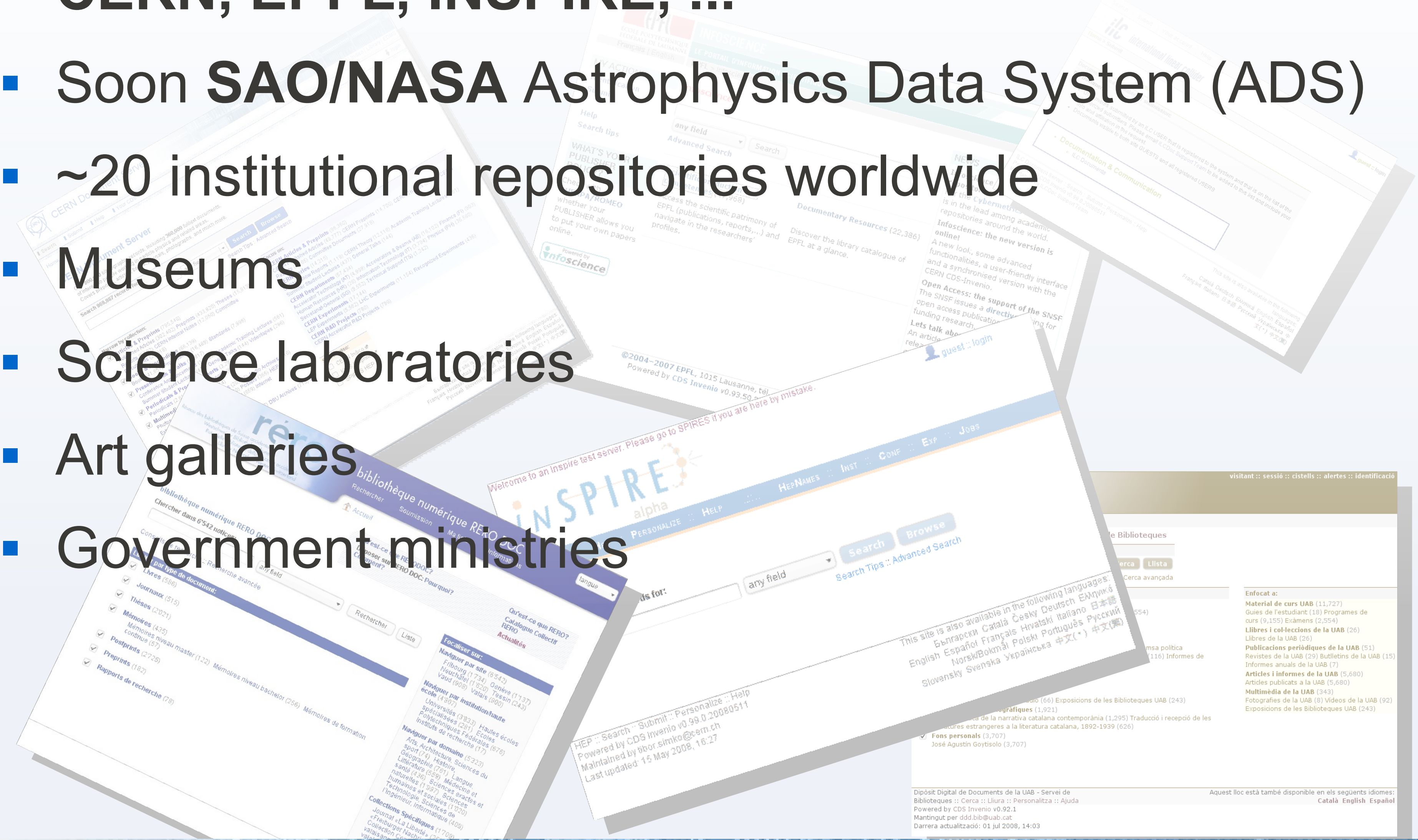
1993	1996	2000	2002	2006
Preprint Server	WebLib	CERN Document Server	CDSware	Invenio

- **2009** – Invenio takes part in EU projects:
  - **OpenAIRE** – to implement the Orphan Record Repository
  - **D4Science II** – through collaboration with INSPIRE
- **2010** – target release date of **Invenio 1.0**

<http://invenio-software.org/>

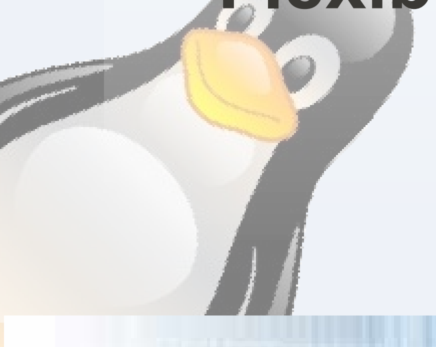
Usage

- **CERN, EPFL, INSPIRE, ...**
- Soon **SAO/NASA** Astrophysics Data System (ADS)
- ~20 institutional repositories worldwide
- Museums
- Science laboratories
- Art galleries
- Government ministries



Architecture

- **Open source** GNU General Public License project
- **Python** (and C and Lisp and Javascript), **MySQL** and **Apache**
- **Modular architecture**
- Implements **open standards** (MARCXML, MARC21, OAI-PMH, OpenURL...)
- **Medium to big** size record repositories (~5M)
- **Flexible** in every layer



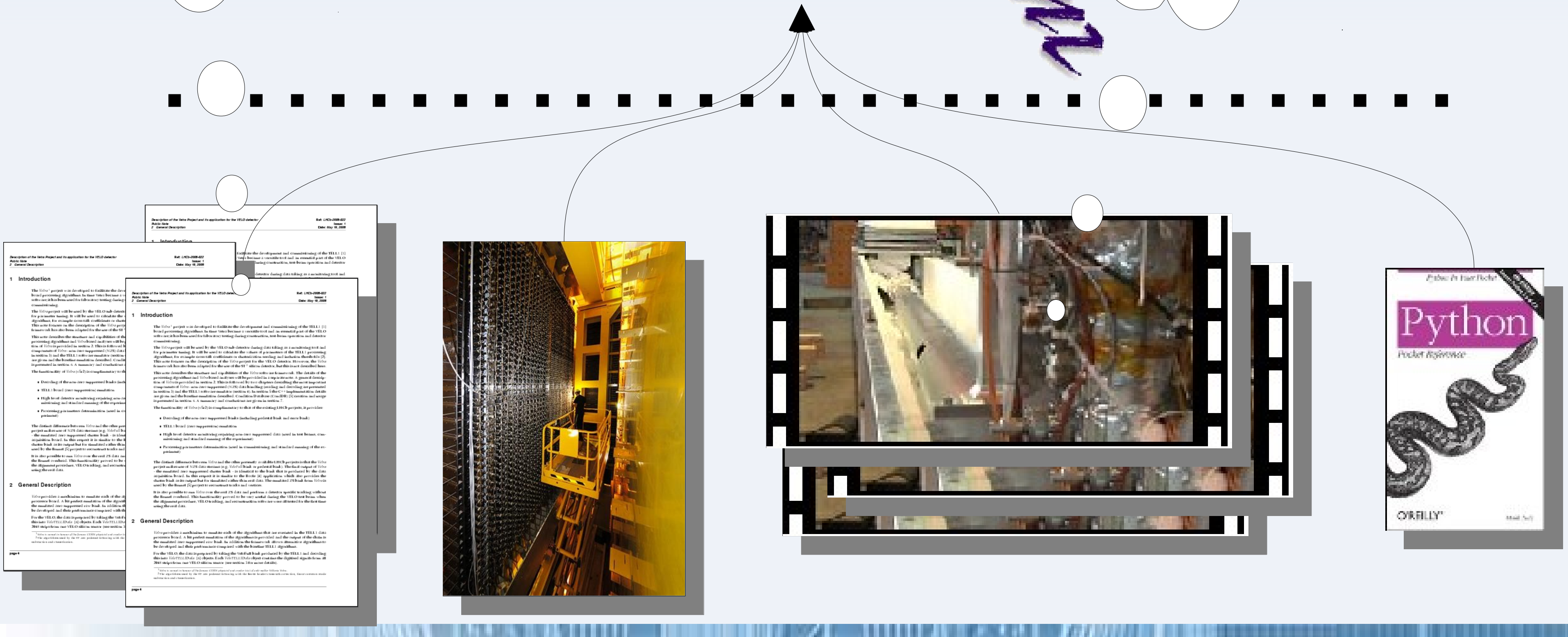
Data and Metadata

Multiple revisions

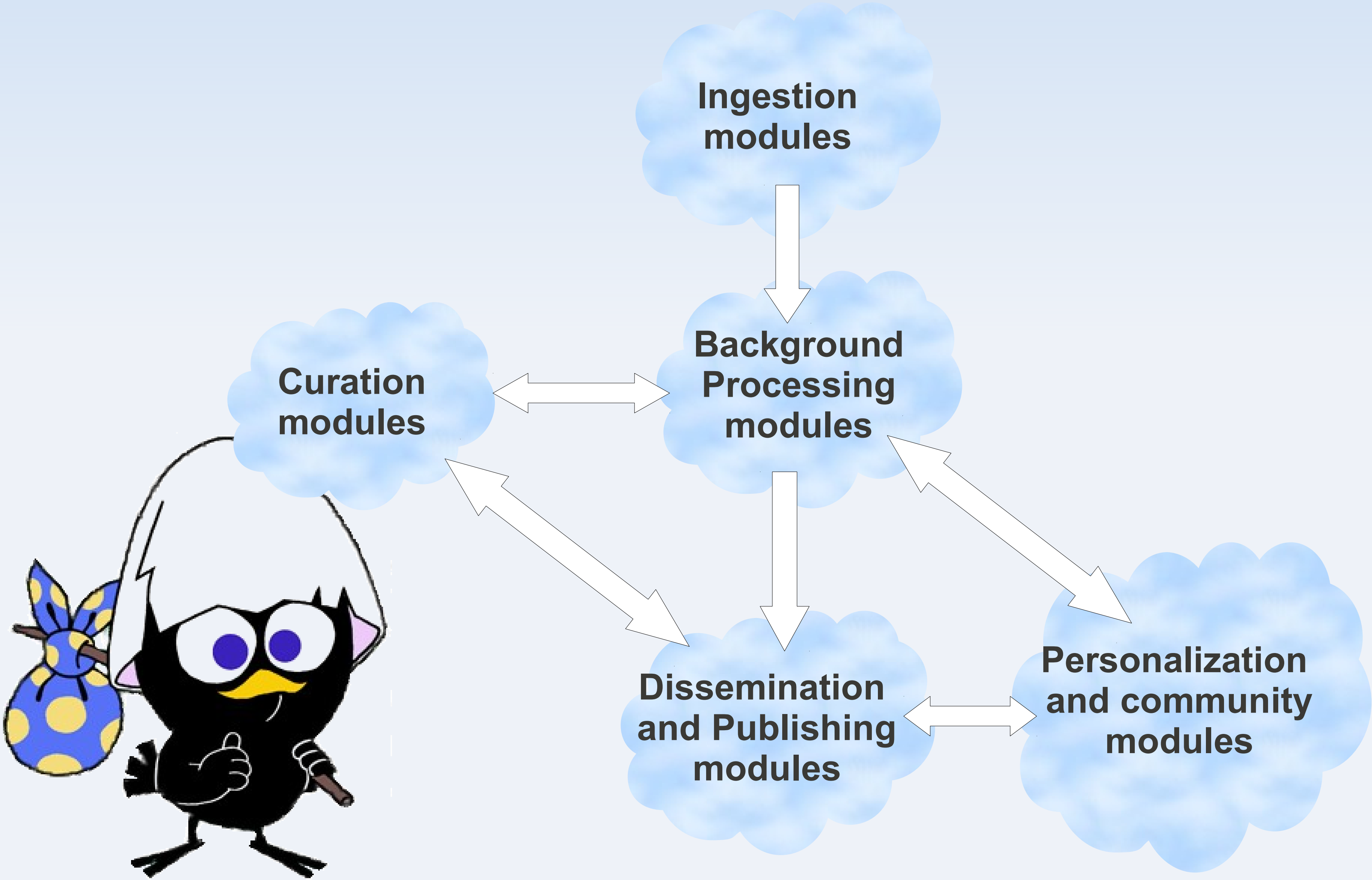
- Who
- When
- Where
- What
- With whom
- ...

marcxml

Multiple formats



The journey of a record


Ingestion modules
Background Processing modules
Curation modules
Dissemination and Publishing modules
Personalization and community modules

Ingestion modules

- **OAI-PMH Harvesting** (DC and MARCXML out of the box)
- **Mail based** document deposition

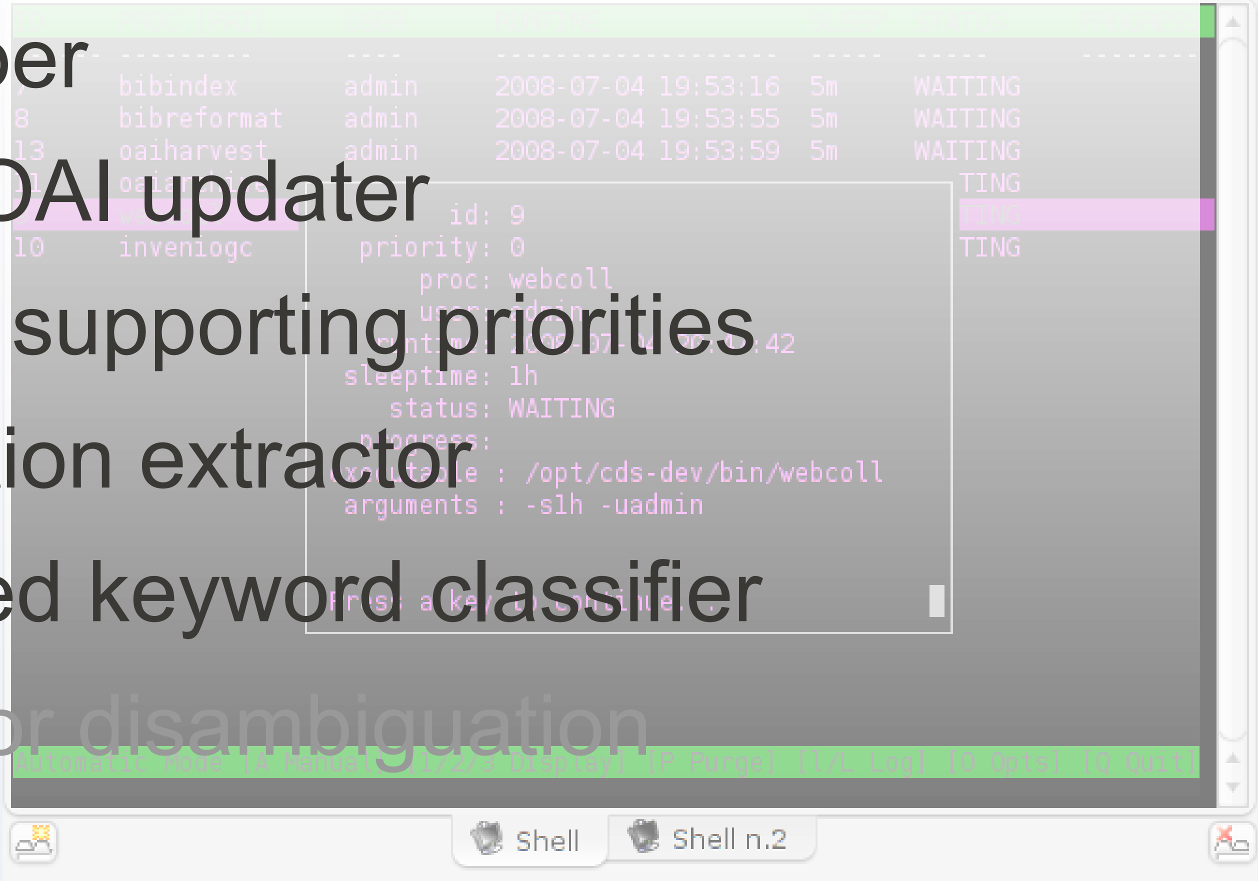
- **Web based** flexible framework to design submission web forms and approval workflows
  - Several workflow building blocks, including fulltext watermarking, plots extractions, ...
- **SWORD** server side (development ongoing)

Background processing modules I

- **Collections** (based on metadata)
- Pre-caching of common **formats**
- Pre-caching of common **splash-pages**
- Variety of **ranking methods**
- Extremely fast **indexing data structure** (to implement IR boolean model)
- Any metadata field can be indexed
- Special treatment for author and journal names, dates and fulltext documents

Background processing modules II

- Garbage collector
- Database dumper
- Harvester and OAI updater
- Task scheduler supporting priorities
- Automated citation extractor
- Taxonomy based keyword classifier
- Automatic author disambiguation

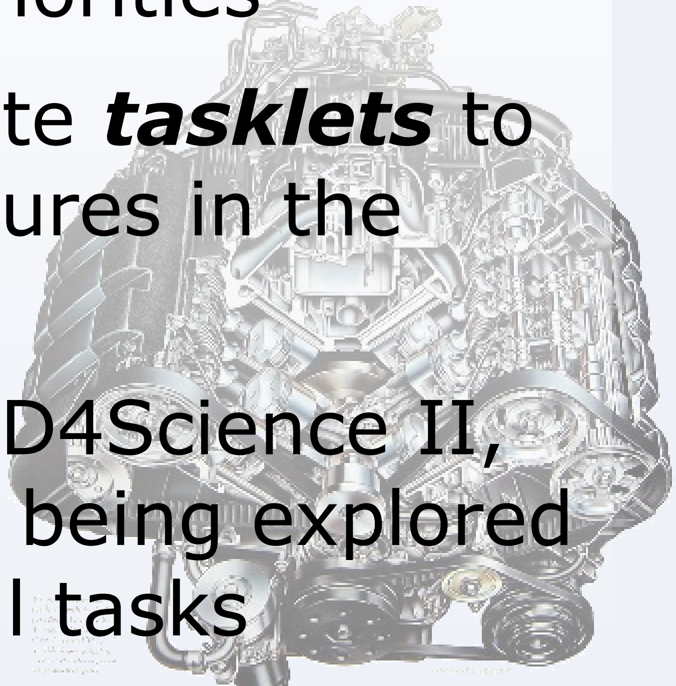


```
7       bibindex      admin    2008-07-04 19:53:16  5m    WAITING
8       bibreformat   admin    2008-07-04 19:53:55  5m    WAITING
13      oaiharvest    admin    2008-07-04 19:53:59  5m    WAITING
11      oaiarchive                                          TING
                                                            TING
10      inveniogc                                           TING
              id: 9
        priority: 0
            proc: webcoll
            user: admin
         runtime: 2008-07-04 [illegible]:42
       sleeptime: 1h
          status: WAITING
        progress:
      executable : /opt/cds-dev/bin/webcoll
      arguments : -slh -uadmin

      Press a key to continue...
Automatic Mode [A Manual] [1/2/3 Display] [P Purge] [l/L Log] [O Opts] [Q Quit]
```

Background processing modules III

- Any Invenio heavy computational task implemented as a *daemon*
- **Bibliographic Task Scheduler** that takes care of executing tasks at proper times, respecting dependencies and priorities
- Plugin framework to quickly write ***tasklets*** to add additional recurrent procedures in the day-to-day repository life
- Thanks to the partnership with D4Science II, cloud and grid computating are being explored to perform heavy computational tasks



Indexing structure

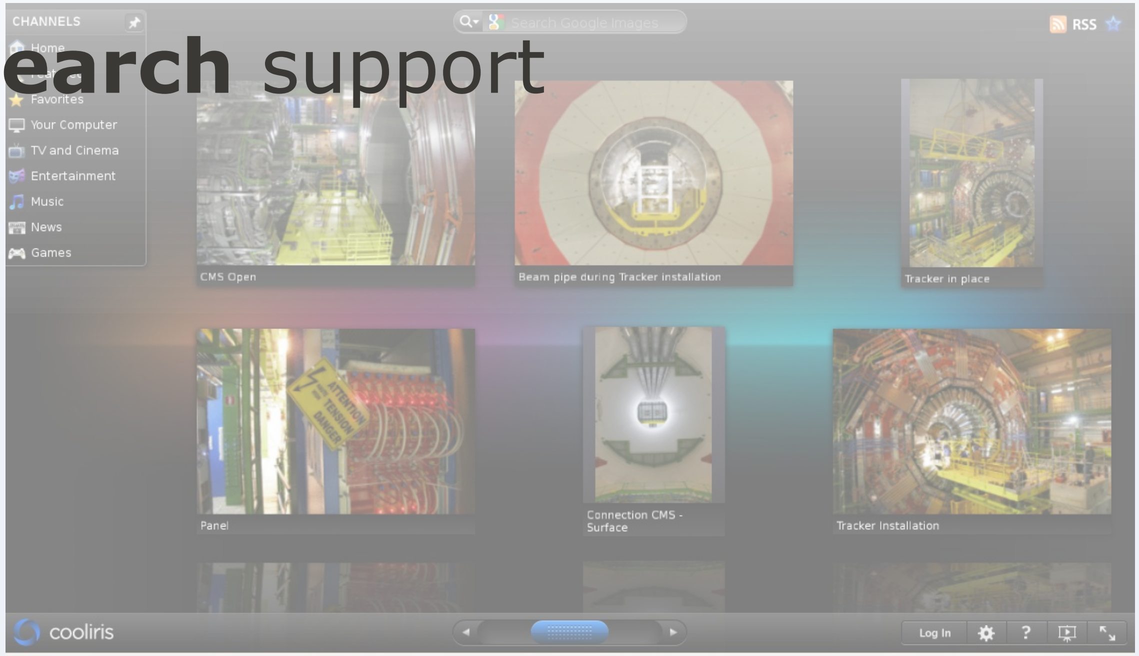
- **Words**, **pairs** of words and **phrase** indexes
- Using forward and reverse indexes
  - By using **extremely fast bit set** in the form of a C extension to Python
- **Stemming**, stop words, TeX formulas, TeX/HTML **stripping**
- Investigation on Solr and Lucene integration
- Planned implementation of generalized *derived logical fields* framework i.e. indexable metadata fields generated on the fly from any piece of information related to a record, via plug-ins

Dissemination and publishing modules I

- Fully configurable **regular and virtual collection tree** (defined on the metadata)
- Multiple pluggable and flexible **output formats** (e.g. MODS, BibTeX, EndNote, Excel, EndNote, DC, NLM, RefWorks...)
- **Fast response** to search queries
- Search available on **any metadata field, fulltext, citations, plot captions...**
- Personalized **RSS feeds**
- **OpenURL** resolver
- **OAI-PMH** 2.0 exporting (MARC and OAIDC out of the box)
- SWORD pushing (currently to arXiv)

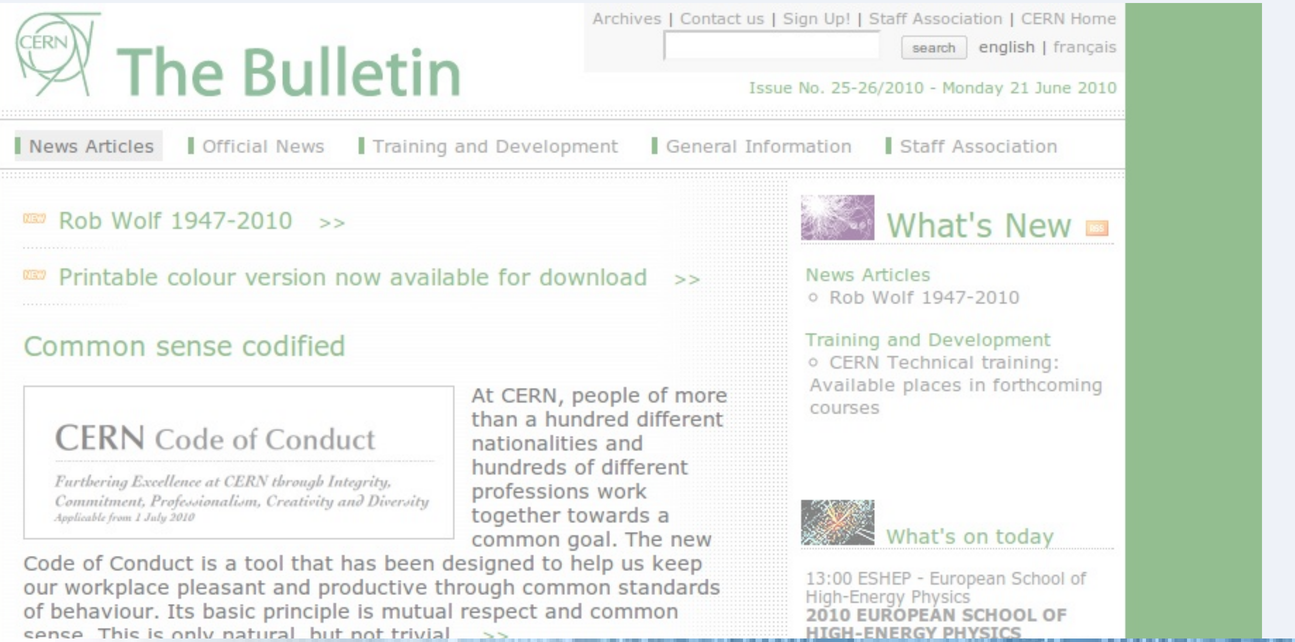
Dissemination and publishing modules II

- Integration with **LibX** Firefox/IE toolbar
- Integration with **Cooliris**
- Integration with **Zotero**
- **OpenSearch** support



Dissemination and publishing modules III

- **Journal** module (web based journal generated from records)
  - An Open Access publishing platform



Archives | Contact us | Sign Up! | Staff Association | CERN Home

search english | français

The Bulletin

Issue No. 25-26/2010 - Monday 21 June 2010

News Articles | Official News | Training and Development | General Information | Staff Association

Rob Wolf 1947-2010 >>

Printable colour version now available for download >>

Common sense codified

CERN Code of Conduct

Furthering Excellence at CERN through Integrity, Commitment, Professionalism, Creativity and Diversity

Applicable from 1 July 2010

At CERN, people of more than a hundred different nationalities and hundreds of different professions work together towards a common goal. The new Code of Conduct is a tool that has been designed to help us keep our workplace pleasant and productive through common standards of behaviour. Its basic principle is mutual respect and common sense. This is only natural, but not trivial >>

What's New

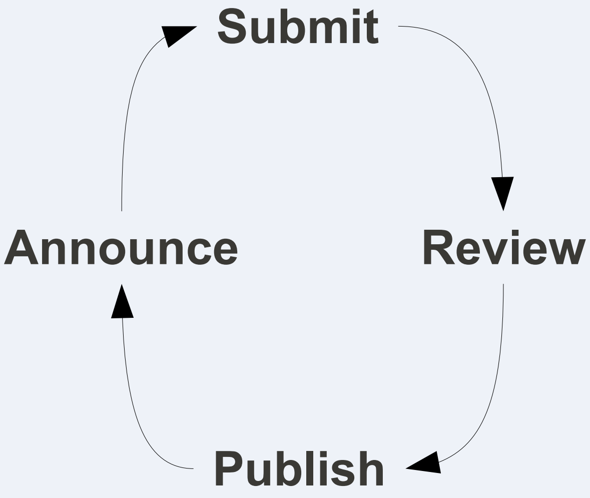
News Articles
- Rob Wolf 1947-2010

Training and Development
- CERN Technical training: Available places in forthcoming courses

What's on today

13:00 ESHEP - European School of High-Energy Physics

2010 EUROPEAN SCHOOL OF HIGH-ENERGY PHYSICS



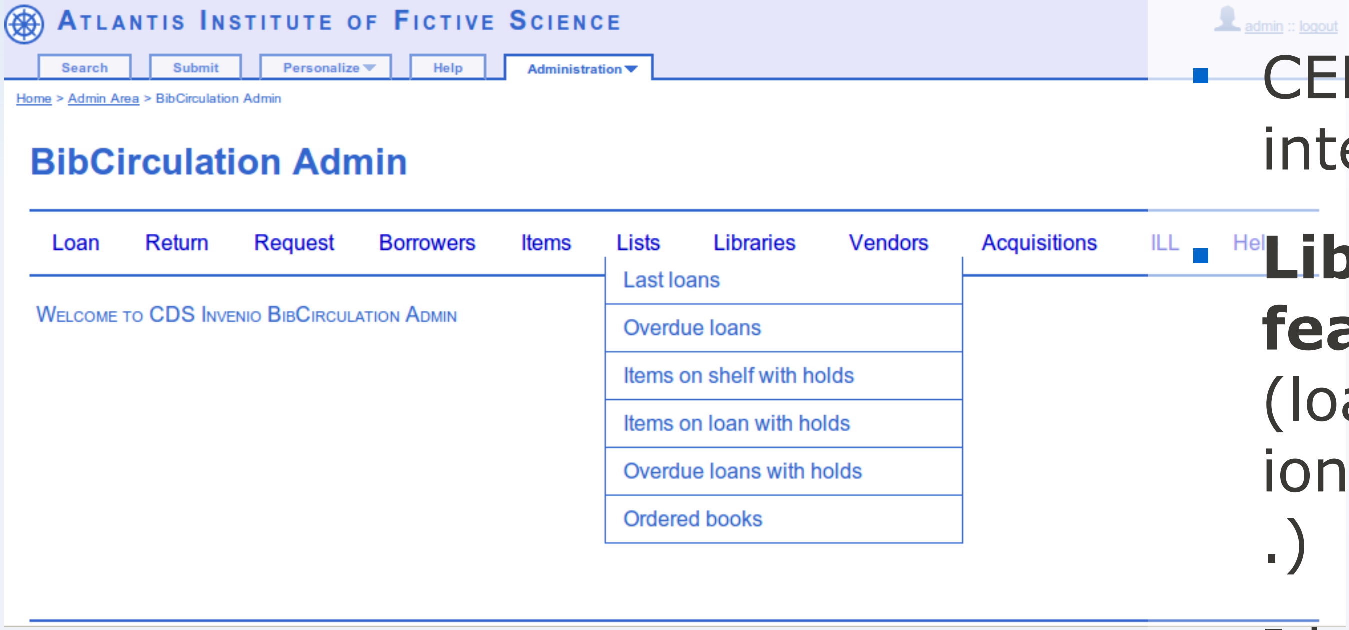
Submit

Review

Publish

Announce

Dissemination and publishing modules IV



BibCirculation Admin

Welcome to CDS Invenio BibCirculation Admin

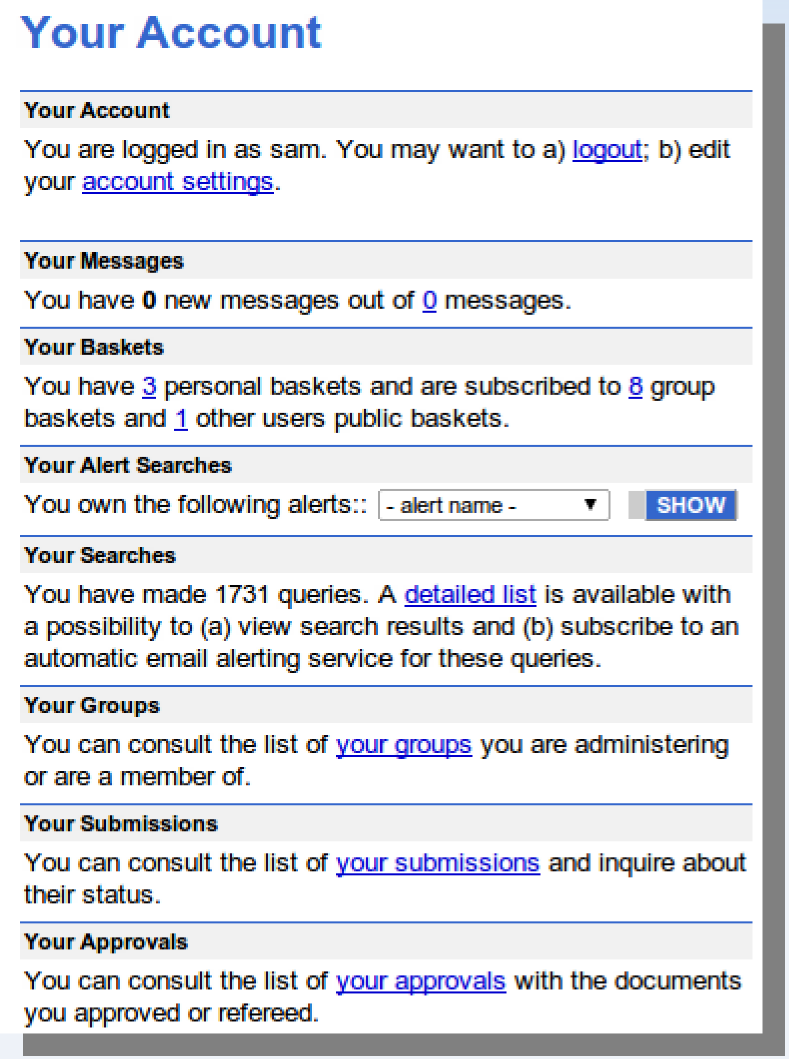
- **Circulation** module (to use Invenio to manage items and borrowers)
- CERN Library catalogue integrated into Invenio
- **Library-specific features** implemented (loans/holdings/reservations/warning/renewals...)
- Ideal solution for "*library meets repository*"

Personalization and community modules I

- **Baskets/bookshelves** (to organize personal collections of interesting records and to share them)
- **Commenting**, **reviewing**
- Web **messages**
- Automatic **alerts** for new results
- **Tagging** (soon to be available)

Personalization and community modules II

- Managing interrupted **submissions**
- Managing **approval** workflows
- Depositing interfaces



Your Account

Your Account

You are logged in as sam. You may want to a) logout; b) edit your account settings.

Your Messages

You have **0** new messages out of 0 messages.

Your Baskets

You have 3 personal baskets and are subscribed to 8 group baskets and 1 other users public baskets.

Your Alert Searches

You own the following alerts:: - alert name - ▼ SHOW

Your Searches

You have made 1731 queries. A detailed list is available with a possibility to (a) view search results and (b) subscribe to an automatic email alerting service for these queries.

Your Groups

You can consult the list of your groups you are administering or are a member of.

Your Submissions

You can consult the list of your submissions and inquire about their status.

Your Approvals

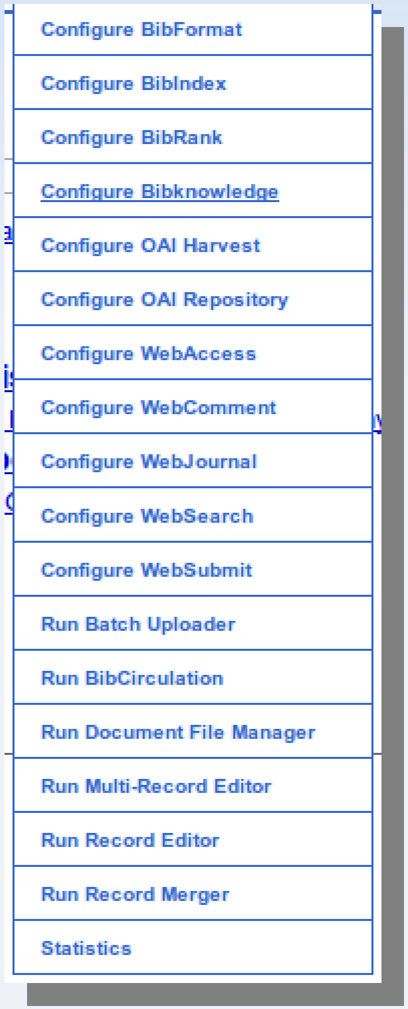
You can consult the list of your approvals with the documents you approved or refereed.

Curation modules I

- **MARC Editor**
  - Ajax record editor, with undo/redo, holding pen, knowledge base lookup, and more
  - Multi-record editor to modify in batch several records
  - Record merger to visualize and merge differences
- **Knowledge bases** (authority files, vocabularies, dynamically generated)
  - Integration with D-NET vocabularies (in the framework of the OpenAIRE collaboration)

Curation modules II

- **Flexible configuration** based on web interfaces for:
  - Logical fields, indexes, ranking algorithms, collection structures, deposition interfaces, collection appearance and hierachy, OAI-PMH harvesting and exporting, formats, ...



Curation modules III

- Automatic metadata **inconsistency checker** (with optional automatic correction)
- Automatic filter of incoming records against current repository to spot e.g. **duplicates**
- **Batch uploading** of fulltext files
- Web interface to **manage full-text files** of a given record
- Author disambiguation tool project

Authentication and Authorization I

- **Plug-in based** authentication framework:
  - To exploit already existing credentials data-bases
  - **LDAP** support
  - **Shibboleth** support to implement Single Sign-On.
- Local and external **groups** (useful for authorization, exchanging messages, defining shared baskets)
- *Role-Based Access Control* (RBAC) authorization model

Authentication and Authorization II

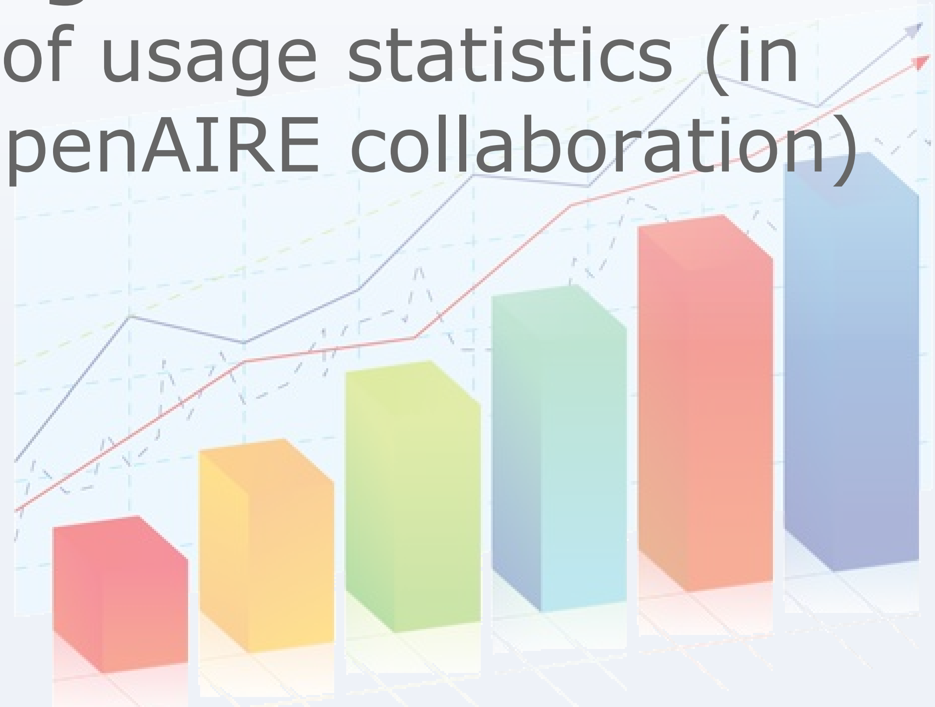
- Role defined:
  - Explicitly by attaching users to roles
  - Implicitly via *Firewall-like Role Definitions*
- **Any user detail** can be used to define a role (i.e. group membership, IP address, regular expression on the institution name...)

Ranking

- Extensible **ranking** framework
- Support out of the box for:
  - **Word similarity** (implementing a flavour of the classical IR vectorial model)
  - advanced **citation network analysis** (co-cited with, time-decayed rankings)
  - **Record access** and **download**

Statistics

- Extensible **event recording** framework
- **Access and download** statistics
- **Knowledge Exchange guidelines** to facilitate the exchange of usage statistics (in the framework of the OpenAIRE collaboration)



Digital objects support I

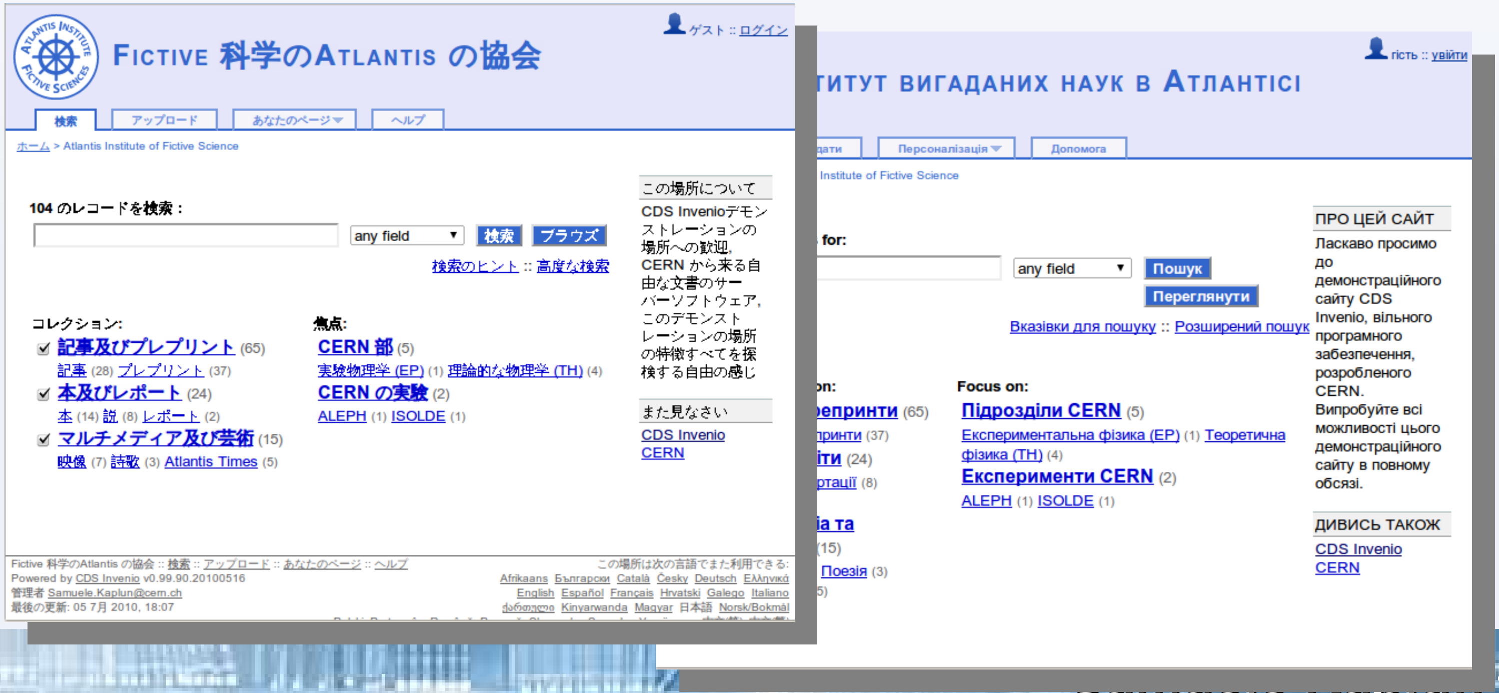
- **Multiple documents** can be attached to a record
- **Multi-format** and **multi-revision** support
- **Icon** creation
- PDF **Stamping**
- **PDF/A** conversion
- Text extraction (including **OCR**)
- **Scanned PDF to PDF** with OCRed text

Digital objects support II

- HTTP streaming supporting **range requests** (e.g. to stream multimedia material, or linearized PDFs)
- Automated **checksum** verification
- **Batch uploading** of documents
- Multiple **protection and authorization** levels
- Automatic and plug-in based **meta-data extraction** (e.g. XMP, EXIV, …)
- Plugin-based **format conversion framework** (development ongoing)

Internationalization

- Invenio interface exists in 26 languages



Fictive 科学のAtlantis の協会
ゲスト :: ログイン
検索
アップロード
あなたのページ
ヘルプ
ホーム > Atlantis Institute of Fictive Science
any field
検索
ブラウズ
検索のヒント :: 高度な検索

104 のレコードを検索：

コレクション:

☑ 記事及びプレプリント (65)
記事 (28) プレプリント (37)
☑ 本及びレポート (24)
本 (14) 説 (8) レポート (2)
☑ マルチメディア及び芸術 (15)
映像 (7) 詩歌 (3) Atlantis Times (5)

焦点:

CERN 部 (5)
実験物理学 (EP) (1) 理論的な物理学 (TH) (4)
CERN の実験 (2)
ALEPH (1) ISOLDE (1)

この場所について

CDS Invenioデモンストレーションの場所への歓迎, CERN から来る自由な文書のサーバーソフトウェア, このデモンストレーションの場所の特徴すべてを探検する自由の感じ

また見なさい

CDS Invenio
CERN

Fictive 科学のAtlantis の協会 :: 検索 :: アップロード :: あなたのページ :: ヘルプ
Powered by CDS Invenio v0.99.90.20100516
管理者 Samuele.Kaplun@cern.ch
最後の更新: 05 7月 2010, 18:07

この場所は次の言語でまた利用できる:
Afrikaans Български Català Česky Deutsch Ελληνικά
English Español Français Hrvatski Galego Italiano
ქართული Kinyarwanda Magyar 日本語 Norsk/Bokmål

титут вигаданих наук в Атлантісі
гість :: увійти
дати
Персоналізація
Допомога
Institute of Fictive Science
any field
Пошук
Переглянути
Вказівки для пошуку :: Розширений пошук

for:

on:

епринти (65)
принти (37)
іти (24)
ртації (8)
ia та
(15)
Поезія (3)
5)

Focus on:

Підрозділи CERN (5)
Експериментальна фізика (EP) (1) Теоретична фізика (TH) (4)
Експерименти CERN (2)
ALEPH (1) ISOLDE (1)

ПРО ЦЕЙ САЙТ

Ласкаво просимо до демонстраційного сайту CDS Invenio, вільного програмного забезпечення, розробленого CERN. Випробуйте всі можливості цього демонстраційного сайту в повному обсязі.

ДИВИСЬ ТАКОЖ

CDS Invenio
CERN

Development support

- ***Hacking guides*** (i.e. recipes on how to extend and tune Invenio even further)
- Unit, regression and web **test suites**
- **Unhandled exceptions** sent via email to admin
  - **SMS sending** for emergencies (e.g. when bibliographic task manager stops after an error)
- Inherent **code quality** checking

Conclusions

- **Extremely flexible** integrated digital library and digital repository software
- Suitable for **medium to large** repositories (up to 10M records)
- **Broad range of features**
- **Highly configurable** to cover any large institution needs

Links

- Invenio home page:
  - <http://invenio-software.org/>
- Invenio demo site (running v0.99.1):
  - <http://invenio-demo.cern.ch/>
  - <http://invenio-software.org/wiki/General/Demo>
- OpenAIRE:
  - <http://www.openaire.eu/>
- INSPIRE:
  - <http://inspirebeta.net/>