OAI6 - Tutorial

Jérôme Caffaro
Samuele Kaplun
1993 - CERN Preprint Server on the Web: institutional repository

1996 - CERN Library Server (weblib): added books and periodicals

2002 - CERN Document Server Software (CDSware): multimedia data and OAI-compliant

2006 – changed name to CDS Invenio

Today - …
1993 - CERN Preprint Server on the Web:

"The Organization shall provide for collaboration among European States in nuclear research of a pure scientific and fundamental character (...). The Organization shall have no concern with work for military requirements and the results of its experimental and theoretical work shall be published or otherwise made generally available".

CERN Convention, Article II.1, 1954
1993 - CERN Preprint Server on the Web: institutional repository

1996 - CERN Library Server (weblib): added books and periodicals

2002 - CERN Document Server Software (CDSware): multimedia data and OAI-compliant

2006 – changed name to CDS Invenio

Today - …
History

1993 - CERN Preprint Server on the Web: institutional repository

1996 - CERN Library Server (weblib): added books and periodicals

2002 - CERN Document Server Software (CDSware): multimedia data and OAI-compliant

2006 – changed name to CDS Invenio

Today - …
What is Invenio?
What is Invenio?
What is Invenio?

User submit

Atlantis Institute of Fictive Science

Search 96 records for:

- any field
- Search
- Browse

Narrow by collection:
- Articles & Preprints (63)
- Books & Reports (24)
- Multimedia & Arts (9)

Focus on:
- CERN Divisions (5)
- CERN Experiments (2)

User
What is Invenio?

User

submit

search
What is Invenio?

User

approve

share

'bookmark'

submit

search

comment, review

subscribe

Invenio is a web-based platform designed for the management and dissemination of scientific information. It allows users to submit, search, comment, review, and share documents. The interface is designed to be user-friendly, with features such as "approve," "share," "'bookmark,'" "submit," "search," "comment, review," and "subscribe."
What is Invenio?

submit

search

'bookmark'

approve

share

comment, review

subscribe

Librarian

edit
What is Invenio?

User:
- submit
- search
- comment
- review
- subscribe

Librarian:
- edit
- harvest
- import
- merge
- convert

approve
share
'bookmark'

What is Invenio?

User:
- submit
- search
- comment, review
- subscribe
- 'bookmark'
- approve
- export
- 'check'
- share

Librarian:
- import, harvest
- merge
- convert
- edit

Admin:
- script
- customize
- monitor

Invenio Interface:

Atlantis Institute of Fictive Science

Submit
Search
Personalize
Help

Search Tips: Advanced Search

Narrow by collection:
- Articles & Preprints (63)
  - Articles (27) Preprints (36)
- Books & Reports (24)
- Multimedia & Arts (9)
- Pictures (1)
- Poetry (1)

Focus on:
- CERN Divisions
  - Experimental Physics (EP)
  - Theoretical Physics (TH)
- CERN Experiments
  - ALEPH
  - DELPHI

GSI searches

This site is also available in the following languages:
- Castellano
- Deutsch
- Espanol
- Français
- Nederlands
- Portugues

Print: List

Samuele Kaplun & Jérôme Caffaro – CERN Document Server team
CERN workshop on Innovations in Scholarly Communication [17th June 2009]
What is Invenio?

Feature-full digital library system
What is Invenio?

Free download at:

http://cdsware.cern.ch/invenio/download.html
What is Invenio?

Free download at:

http://cdsware.cern.ch/invenio/download.html

Sources (Git):

http://cdsware.cern.ch/repo/cds-invenio.git
OpenSource GPL project
Python (and C and Lisp), MySQL and Apache
Modularized architecture
Based on open standards (MARCXML, MARC21, OAI-PMH, OpenURL...)
Medium to big data repositories
Flexible in every layer
How to install

- Unix/Linux
- Python
- Apache + mod_python
- MySQL + MySQLdb
- Other smaller dependencies

See: [http://cdsware.cern.ch/download/INSTALL](http://cdsware.cern.ch/download/INSTALL)

Contact [project-cdsware-users@cern.ch](mailto:project-cdsware-users@cern.ch) for help
Workflow
Workflow

OAI Data Source  Non-OAI Data Source  Author
OAI  mail  web interface

BibHarvest  ElmSubmit  WebSubmit
BibCheck  BibConvert  BibUpload
BibEdit

Library Cataloguer

Clusters  Formats  Rankings  Indexes
BibSched  BibFormat  BibRank  BibIndex

BibSched

WebColl  Bibliographic Metadata Server
WebSearch

WebDoc  WebHelp

WebMessage

OAI Service Provider

OAI  mail  web interface

WebSession  WebAccess

Reader  Group  Anybody

FullText
30 modules
~200,000 Python Lines of code
~500 test cases
Workflow
Jérôme Caffaro & Samuele Kaplun – CERN Document Server team
CERN workshop on Innovations in Scholarly Communication [17th June 2009]
Workflow

OAI Data Source

OAI

Non-OAI Data Source

mail

BibHarvest

ElmSubmit

BibCheck

BibEdit

WebColl

BibFormat

BibRank

BibIndex

WebSearch

BibHarvest

WebAlert

WebBasket

WebComment

WebStat

WebMessage

Reader

Group

Anybody

Fulltext Server

BibSched

Bibliographic Metadata Server

BibConvert

BibUpload

BibSched

Fulltext Document Server

BibClassify

RefExtract

WebSubmit

web interface

Meta Data

Full Text

web message

WebAccess

WebSession

Anybody

OAI Data Source

OAI Service Provider

Jérôme Caffaro & Samuele Kaplun – CERN Document Server team
CERN workshop on Innovations in Scholarly Communication [17th June 2009]
Clusters Formats Rankings Indexes

OAI Data Source Non-OAI Data Source

OAI

BibHarvest ElmSubmit

BibCheck BibEdit

Library Cataloguer

Author

WebAccess WebSession

web interface

WebSubmit

BibConvert

BibUpload

BibSched

BibCheck BibEdit

Library Cataloguer

Fulltext Document Server

BibSched

WebColl BibFormat BibRank BibIndex

Bibliographic Metadata Server

WebConvert

WebSubmit

WebHarvest

WebAlert WebBasket WebComment WebStat

web interface

Reader Group Anybody

WebMessage

OAI Service Provider

OAI

mail

web interface

OAI

mail

web interface

WebMessage

WebDoc WebHelp

WebSearch
Workflow
Workflow
BibSched

Background daemons
- Inputting
- Indexing
- Pre-caching
- Monitoring
- Tasks scheduling
Integration & Customization
Jérôme Caffaro & Samuele Kaplun – CERN Document Server team

CERN workshop on Innovations in Scholarly Communication [17th June 2009]
Style Customization

Collection tree
Style Customization

Portal boxes

Collection tree
Style Customization

Each building block of the web interface

Portal boxes

Collection tree

Templating system, with WebStyle
Style Customization

- Templating system, with WebStyle
Style Customization

- Templating system, with WebStyle
Style Customization

- Templating system, with WebStyle
Style Customization

- Templating system, with WebStyle
Collection Tree
Collection Tree

Regular collections
("narrow by")

Virtual collections
("focus on")
Submissions

- Customizable web forms
- Customizable workflow (approval process, sending emails, conversion, stamping, OCR, etc)
- Metadata & File revisions preserved
Customizable web forms

Customizable workflow (approval process, sending emails, conversion, stamping, OCR, etc)

Metadata & File revisions preserved
Record formatting

- HTML-like
- XSL
Abstract:

DIRAC is the LHCb gateway to any computing grid infrastructure (currently supporting WLCG) and is intended to reliably run large data mining activities. The DIRAC system consists of various services (which wait to be contacted to perform actions) and agents (which carry out periodic activities) to direct jobs as required. An important part of ensuring the reliability of the infrastructure is the monitoring and logging of these DIRAC distributed systems.
Authentication

To support an external authentication method (e.g. to interact with an institutional Human Resources Database, implementing):

- user/password checking
- External group/mailing list membership importing
- External user details (e.g. phone number, affiliation...)

**Authentication methods:**

- «Internal»
- Shibboleth
- LDAP
- ... (write your plugin)
Access Control

- RBAC (Role Based Access Control)
  http://invenio-demo.cern.ch/help/admin/webaccess-admin-guide

- FireRole (Firewall Like Role Definition)
  http://invenio-demo.cern.ch/help/admin/webaccess-admin-guide#6
Access Control

- RBAC (Role Based Access Control)
  http://invenio-demo.cern.ch/help/admin/webaccess-admin-guide

- FireRole (Firewall Like Role Definition)
  http://invenio-demo.cern.ch/help/admin/webaccess-admin-guide#6

allow email /.cern.ch/,/.*@slac.stanford.edu/
deny group badguys
allow remote_ip "123.0.0.0/24"
deny all
Tools (Cataloguing)
Record enrichment

- BibMerge to merge records
- BibCheck batch record checker and corrector
- RefExtract to extract reference lists from PDFs
- BibClassify to extract keywords (RDF taxonomy)
Looking for help?

e.g. [http://invenio-demo.cern.ch/help/admin/](http://invenio-demo.cern.ch/help/admin/)

<table>
<thead>
<tr>
<th>Admin Module</th>
<th>Admin Description</th>
<th>Admin Interface</th>
<th>Admin Guide</th>
</tr>
</thead>
<tbody>
<tr>
<td>BibIndex Admin</td>
<td>Enables you to configure &quot;word files&quot;, i.e. to define which bibliographic fields are indexed into which word indexes. The word indexes are then used by the search interface. For example, you can define that the logical author index is constructed from physical 100 $a$ and 700 $a$ bibliographic tags, you can force reindexing of the fulltext index, etc.</td>
<td>Manage indexes</td>
<td>BibIndex Admin Guide</td>
</tr>
<tr>
<td>BibRank Admin</td>
<td>Enables you to configure various ranking methods to be used by the search engine. You can rebalance existing ranking sets, create new ranking methods, etc.</td>
<td>BibRank Admin Interface</td>
<td>BibRank Admin Guide</td>
</tr>
<tr>
<td>BibClassify Admin</td>
<td>Enables you to automatically classify documents according to keyword taxonomies and thesauri.</td>
<td>command-line configuration</td>
<td>BibClassify Admin Guide</td>
</tr>
<tr>
<td>BibFormat Admin</td>
<td>Enables you to specify how the bibliographic data is presented to the end user in the search interface. You can decide that titles should be presented in bold font, that for each author an automatic link to author’s home page should be created according to some receipt, etc.</td>
<td>BibFormat Admin Interface</td>
<td>BibFormat Admin Guide</td>
</tr>
<tr>
<td>WebSearch Admin</td>
<td>Enables you to configure the search interface for various metadata collections. You can define new collections and organize them in the tree, you can define various portalboxes that would appear on the screen, you can define search options and search fields to present, etc.</td>
<td>WebSearch Admin Interface</td>
<td>WebSearch Admin Guide</td>
</tr>
<tr>
<td>WebStat Admin</td>
<td>Enables you to configure the usage statistics reporting system.</td>
<td>command-line configuration</td>
<td>WebStat Admin Guide</td>
</tr>
</tbody>
</table>
CERN workshop on Innovations in Scholarly Communication [17th June 2009]

CDS Invenio homepage: http://cdsware.cern.ch/invenio/
CERN Document Server: http://cdsweb.cern.ch/
CDS Invenio Demo installation: http://invenio-demo.cern.ch/
Administrating CDS Invenio: http://invenio-demo.cern.ch/help/admin/
Hacking CDS Invenio: http://invenio-demo.cern.ch/help/hacking/
CDS Invenio Wiki: http://twiki.cern.ch/twiki/bin/view/CDS/Invenio
Thank You!