Managing an Institutional Repository with CERN Document Server Digital Library Software

Diane Berkovits, Martin Vesely

ECDL, 21 September 2006
Term clarification

- **CERN**: European Organization for Nuclear Research
- **CERN Document Server (CDS)**: CERN Institutional repository powered by Invenio
- **CDS Invenio**: in-house developed CERN Document Server Digital Library Software
- **CERN Document Server Software (CDSWare)** is now **CDS Invenio** and **CDS Indico** (Conference management system)
Overview

■ CERN Document Server (CDS)
  → How CERN manages its scholarly literature

■ Invenio
  → Presentation of the application behind CDS

■ R&D
  → On-going research and development at Invenio
CDS history

Number of records and fulltext files per year

- records
- full-text

1954 CERN Creation
1950
1960
1970
1980
1990
2000
2005

Dissemination channels

Library paper preprints distribution
CERN Preprints Server
CDS

LEP is closed down for building LHC machine
LEP machine starts up
Major discoveries made at CERN
First major machine in operation
WWW

D. Berkovits et M. Vesely / Managing an Institutional Repository with CERN Document Server Digital Library Software / ECDL 2006 – p. 4
CDS numbers

- CDS contains today:
  - 850,000 bibliographic records organised in 550 collections
  - 450,000 fulltext documents
  - 60,000 new acquisitions per year (2,000 produced at CERN)

- CDS is used today by:
  - 20,000 unique visitors per month → 80% non-CERN users
  - 200,000 searches per month (≈9,000 queries per day)
CDS: Open Access (1)

OA Service Provider for High Energy Physics (HEP):

♦ Collecting records from more than 100 sources (including ArXiv.org)
♦ Harvesting via OAI-PMH, email/database alerts or web page
♦ Records enriched by CERN Library and Invenio processes
♦ Entirely searchable via web portal (HTML, DC, MARC21, etc)
CDS: Open Access (2)

■ OA Data Provider of CERN production:
  ♦ CDS re-exposes (via OAI) only records for CERN-authored documents
  ♦ Collecting documents via email and web submission by authors
  ♦ 76,000 CERN documents organised in OAI sets (monthly increase of 1,000 records)

→ OA features all supported with Invenio
Invenio: General information

- Open Archive Initiative Protocol (OAI-PMH) compliant: harvesting and serving of records
- Free software (GNU-GPL)
- Uses standard metadata format (MARC21) as its underlying format
- Powered by Python and MySQL
- Multilanguages interface (>15) and user personalisation
- Consortium with EPFL (Swiss Federal Institute of Technology Lausanne) and other contributors
Invenio: Main features

■ Specially designed to cope with very large repositories
■ All modules provide web-based administrative toolkit
■ Designed indexes to provide powerful search engine with Google-like syntax
■ Flexible document-type submission through different channels
■ Customisable interface and metadata representation
■ Metadata organised into several navigable collections
Object and Collection Management

CERN Document Server

Over 800,000 bibliographic records, including 360,000 fulltext documents, of interest to people working in particle physics and related areas. Covers preprints, articles, books, journals, photographs, and much more.

Search 872,480 records for:

Search Tips: Advanced Search

Virtual collections [subject based navigation]

Regular collections [document type navigation]

Narrow by collection:

✓ Articles & Preprints (719,450)
  Published Articles (273,997) Preprints (361,376) Theses (42,807)
  Reports (5,450) CERN Internal Notes (10,732) Committee Documents (26,751)

✓ Books & Proceedings (57,547)
  Books (33,525) Proceedings (16,472) Standards (7,550)
Invenio: Modules overview
R&D: Keyword extraction

- Automatic extraction of most relevant keywords from a controlled vocabulary (thesaurus)
R&D: Ranking

**Goal:** to increase the ranking of positive results in a scholarly literature search

**Methods:**
- Word similarity
- Number of downloads
- Number of citations
- Journal Impact Factor

**Research ongoing:**
- Distributed Ranking method: **DRank**
- Combining ranking methods
- Aim to be find a standard way to rank across different collection types
R&D: Usage Statistic

Downloads history and user repartition

People who viewed this page also viewed
(1) record ID 1
(3) Beam transfer functions and beam stabilisation in a double RF system - Shaposhnikova, E at al. - CERN/LEP 2005/016

Cited by: 2333 records
(943) An Alternative to Compactification - Randall, L et al. - hep-th/9902064
(48) Anti-de Sitter Space, Thermal Phase Transition, And Confinement in Gauge Theories - Witten, Edward - hep-th/9803131
(37) Strings in flat space and pp waves from $S$ (cal N) = 4 $S$ Super Yang Mills - Berenstein, D E et al. - hep-th/0202021
(28) Supergravity and The Large $S$ $S$ Limit of Theories With Sixteen Supercharges - Itzhaki, N et al. - hep-th/9802042
more

Citation history

Graph
R&D: Collaborative tools

- User-based digital library:
  - Comments and reviews: content evaluation
  - Baskets: personal collections of documents
  - Alerts: automatic notifications via email
    → Owned/shared between users/groups or public

- Personalisation of Digital Library System

- Allow users to collaborate in more effective ways

- Combined with the **Search Engine** → Increase work efficiency!
WebBasket

Programming (1)

Addiction to C
2 records - last update: 13 Sep 2006, 20:26

Tim Berners-Lee
28 Jun 1998

Abstract: Conference "Internet, Web, What's next?" on 26 June 1998 at CERN: Tim Berners-Lee, inventor of the World-Wide Web and Director of the W3C, explains how the Web came to be and give his views on the future. [...]

Reviews and comments

People who viewed this page also viewed:

Rate this document:
Average review score: ★★★★★ based on 1 reviews
Readers found the following reviews to be most helpful.

★★★★★ Test review
Reviewed by admin on 14 Sep 2006, 15:12
0 out of 0 people found this review useful
This is a test for ECDL 2006

Report abuse
Was this review helpful? Yes / No

View all 1 reviews

Write a review

Discuss this document:
Start a discussion about any aspect of this document.

Write a comment
Conclusion

■ **CDS Invenio**: a powerful, flexible solution suitable to handle very large collections of full text documents

■ Certified at CERN as the long term electronic archive of our institution

■ Tries to combine the best of the traditional library world and modern information retrieval technology

■ Designed to cope with new dissemination channels of the scientific results of LHC (Open Access)
Thanks and good bye! Questions?

http://cdsware.cern.ch

http://cdsweb.cern.ch

cds.support@cern.ch