Introduction to CDS-Invenio Digital Library Software

Jean-Yves Le Meur - CERN Document Server Team

September 20, 2009

Since the creation of CERN 55 years ago, library mission is the same: the dissemination and long term keeping of High Energy Physics results. Only the means have changed:

From paper to digital age:

Invenio Overview

- Suite of applications and tools enabling the operation and maintenance of: electronic preprint server, digital library catalogue or document archive
- Used to manage CERNs institutional scientific repository:
  - About 1 million records; 550 collections; 10,000 searches/day
  - Wide range of content: documents (articles, preprints, etc), multimedia (photos, videos) and more
  - Designed to cope with new dissemination channels of scientific results of LHC (Open Access)
- Makes use of existing standards:
  - US Library of Congress standards for bibliographic information description (MARC 21, MARCXML)
  - Unicode
  - OAI-PMH

Licensing

- Open source: GNU GPL License
- Regular public releases of software packages
- Support modes
  - Free via listboxes
  - Charged
- CDSware Development Consortium
  - Main partners: EPFL, EIF; exchanging students, code, strategy
  - HEP Collaboration: SLAC, DESY, FERMILAB (INSPIRE Project)
  - World wide contributions; internationalization
  - Open to newcomers !

Prehistory

- 1954 - paper dissemination by the CERN library

Prehistory

- 1954 - paper dissemination by the CERN library
- 1993 - CERN Preprint Server on the Web: institutional repository.

Prehistory

- 1954 - paper dissemination by the CERN library
- 1993 - CERN Preprint Server on the Web: institutional repository.
- 1996 - CERN Library Server (weblib): added books , periodicals and library objects.

Prehistory

- 1954 - paper dissemination by the CERN library
- 1993 - CERN Preprint Server on the Web: institutional repository.
- 1996 - CERN Library Server (weblib): added books , periodicals and library objects.
- 1999 - CERN Agenda (sister application for conferences, meetings, lectures)

Prehistory

- 1954 - paper dissemination by the CERN library
- 1993 - CERN Preprint Server on the Web: institutional repository.
- 1996 - CERN Library Server (weblib): added books , periodicals and library objects.
- 1999 - CERN Agenda (sister application for conferences, meetings, lectures)
- 2000 - CERN Document Server (added multimedia material, internal notes)

History

- 2002 - first release of CERN Document Server digital library software (CDSware), OAI-compliant.

History

- 2002 - first release of CERN Document Server digital library software (CDSware), OAI-compliant.
- 2002 - CDSware starts to be distributed worldwide:
  - SDSC, San Diego, USA; HBZ NRW, Cologne, Germany; Aristotle University of Thessaloniki, Greece and many more...

History

- 2002 - first release of CERN Document Server digital library software (CDSware), OAI-compliant.
- 2002 - CDSware starts to be distributed worldwide:
  - SDSC, San Diego, USA; HBZ NRW, Cologne, Germany; Aristotle University of Thessaloniki, Greece and many more...
- 2002 - start of EU project InDiCo (successor to CDS Agenda)

History

- 2002 - first release of CERN Document Server digital library software (CDSware), OAI-compliant.
- 2002 - CDSware starts to be distributed worldwide:
  - SDSC, San Diego, USA; HBZ NRW, Cologne, Germany; Aristotle University of Thessaloniki, Greece and many more...
- 2002 - start of EU project InDiCo (successor to CDS Agenda)
- 2004 - first release of CDS Indico

History

- 2002 - first release of CERN Document Server digital library software (CDSware), OAI-compliant.
- 2002 - CDSware starts to be distributed worldwide:
  - SDSC, San Diego, USA; HBZ NRW, Cologne, Germany; Aristotle University of Thessaloniki, Greece and many more...
- 2002 - start of EU project InDiCo (successor to CDS Agenda)
- 2004 - first release of CDS Indico
- 2004 - CERN-EPFL collaboration on CDSware partnership and co-development

History

- 2002 - first release of CERN Document Server digital library software (CDSware), OAI-compliant.
- 2002 - CDSware starts to be distributed worldwide:
  - SDSC, San Diego, USA; HBZ NRW, Cologne, Germany; Aristotle University of Thessaloniki, Greece and many more...
- 2002 - start of EU project InDiCo (successor to CDS Agenda)
- 2004 - first release of CDS Indico
- 2004 - CERN-EPFL collaboration on CDSware partnership and co-development
- 2006 CDSware becomes CDS Invenio

Modern History

- 2007 - started collaboration with SPIRES
  - INSPIRE = Invenio + SPIRES (projected production in 2010)
  - world-wide collaboration: CERN, DESY, Fermilab, SLAC
  - aims to create a new single-stop shop for HEP documents

Modern History

- 2007 - started collaboration with SPIRES
  - INSPIRE = Invenio + SPIRES (projected production in 2010)
  - world-wide collaboration: CERN, DESY, Fermilab, SLAC
  - aims to create a new single-stop shop for HEP documents
- 2008 - work on cataloging functionality
  - advanced metadata maintenance tools
  - advanced metadata enrichment tools
  - citation analysis and keyword ontology tools
  - author disambiguation tools

Modern History

- 2007 - started collaboration with SPIRES
  - INSPIRE = Invenio + SPIRES (projected production in 2010)
  - world-wide collaboration: CERN, DESY, Fermilab, SLAC
  - aims to create a new single-stop shop for HEP documents
- 2008 - work on cataloging functionality
  - advanced metadata maintenance tools
  - advanced metadata enrichment tools
  - citation analysis and keyword ontology tools
  - author disambiguation tools
- 2009 - projected release of Invenio 1.0

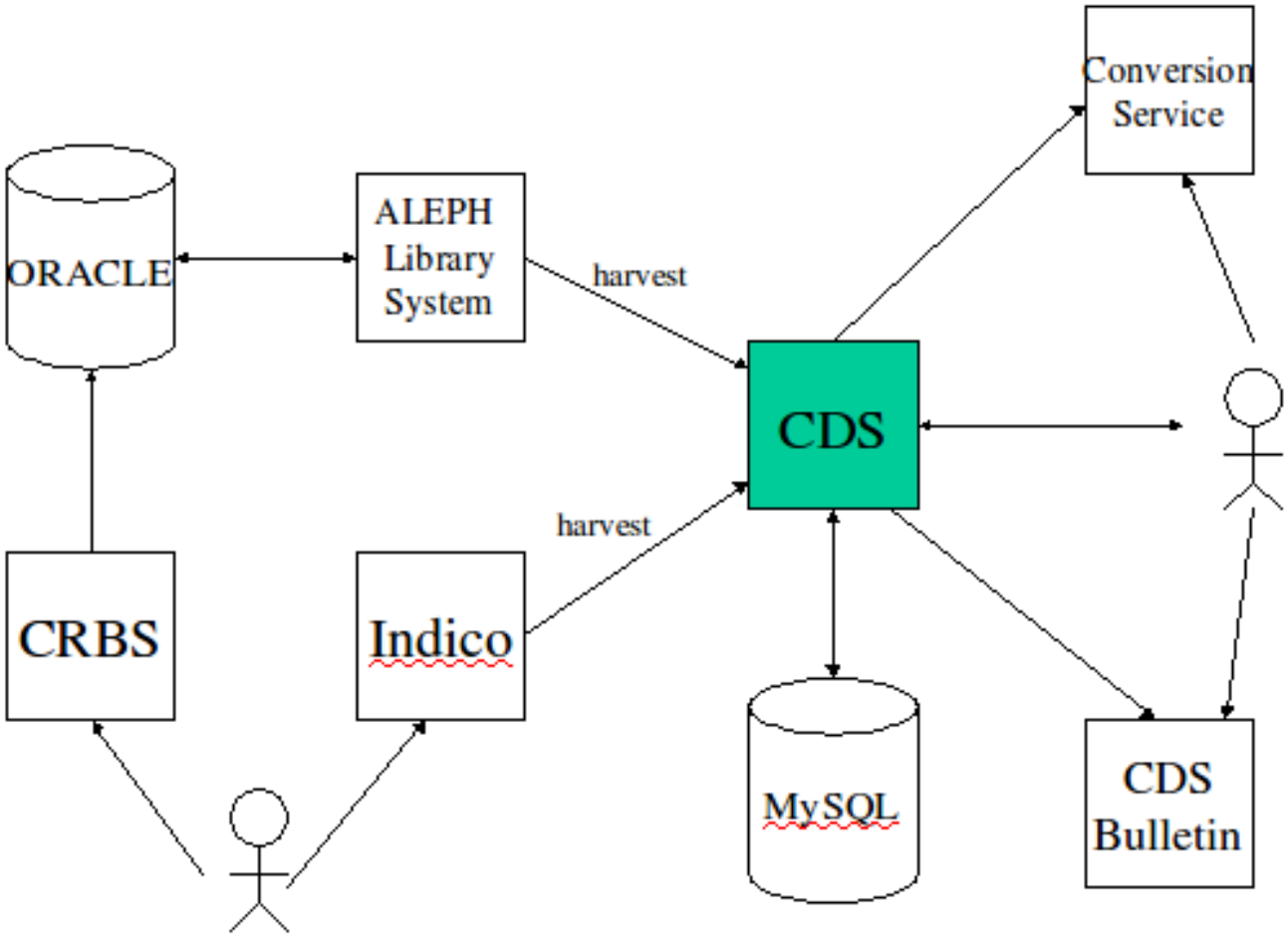
Invenio and other systems I

- Invenio and Library Automated Systems
  - traditional library systems (Aleph500, VTLS, etc) manage books, periodicals, series but do not cope well with heterogeneous material and grey literature
  - specific library-oriented features and use experience
  - strong focus on metadata management, not on full text
- Invenio and EDMS Systems (GED)
  - very different scope & architecture
  - specific features linked to equipment/component management
  - long term archive versus working environment

Invenio and other systems II

- Invenio and Web Search Engines
  - Invenio is a collector to provide access to the "invisible" web
  - full text indexing but poor metadata
  - acquisition process based on harvesting only
  - very limited collection organisation
- Invenio and Institutional Repository Software
  - similar products running institutional repositories: DSpace, EPrints, Fedora...
  - support small repositories only (less than 50K documents) with simple requirements
  - not library-oriented (limited metadata & acquisition channels)

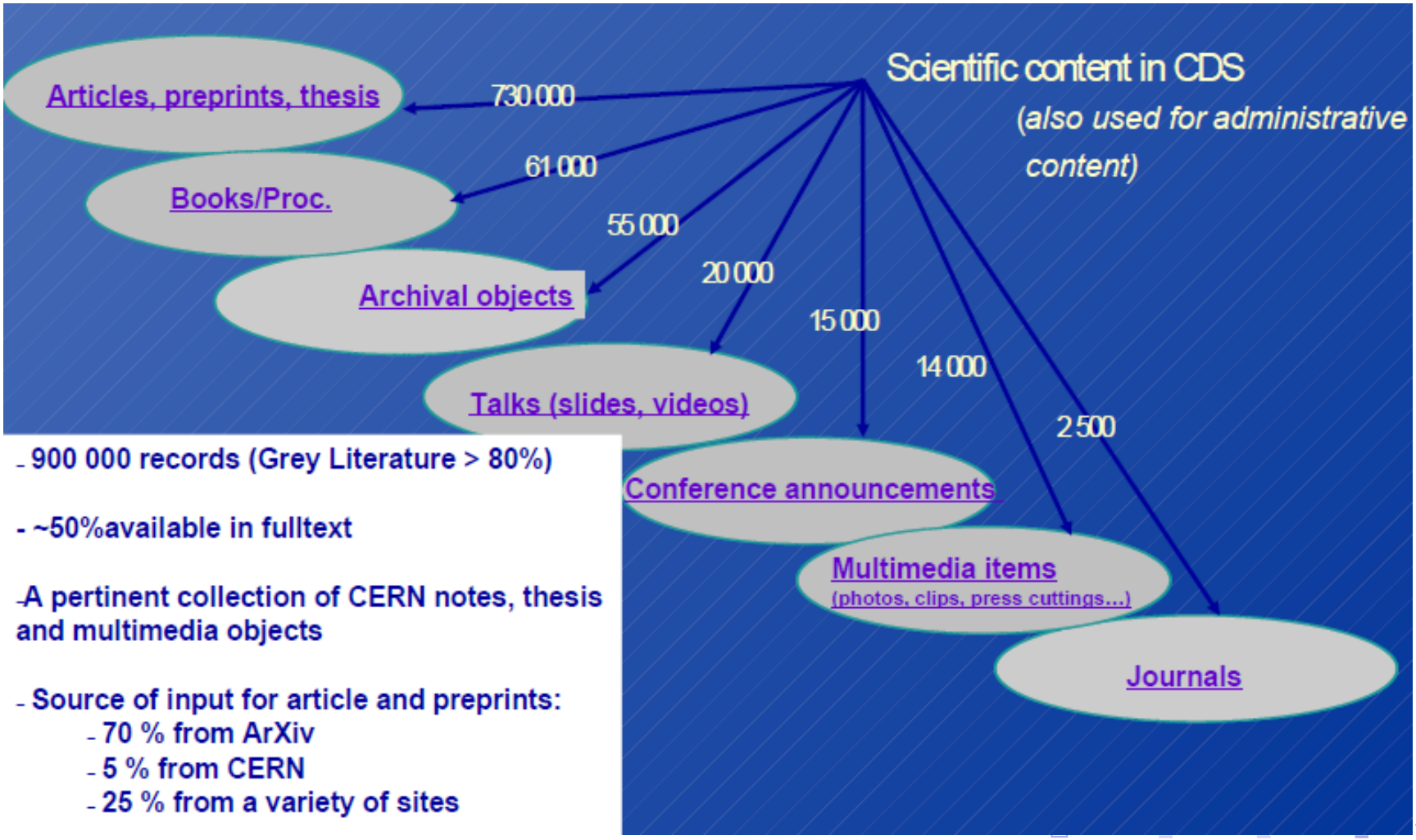
Invenio at CERN: systems



Invenio at CERN: scope

- CERN official repository (e-archive) but also document management system (for directorate offices)
- Search engine used as a back-end platform for Web front-end applications
  - Electronic Bulletins: e.g. http://bulletin.cern.ch
  - Generation of Lists (publications, events, etc)
  - Conference and meetings (Indico) search
- Additional applications running alongside Invenio
  - Document format conversion: CERN Conversion Server
  - Multimedia conversion and analysis (R&D)

Invenio at CERN: scope



- 900 000 records (Grey Literature > 80%)
- ~50%available in fulltext
- A pertinent collection of CERN notes, thesis and multimedia objects
- Source of input for article and preprints:
  - 70 % from ArXiv
  - 5 % from CERN
  - 25 % from a variety of sites

Invenio at CERN: content

- 600 collections; variety of document types (books, preprints, photos, audio, video, . . .)
- 950,000 bibliographic records, 450,000 fulltext files
- 60,000 new acquisitions per year (2,000 by CERN authors)
- institutional repository and subject-based repository search engine for Indico (450,000 meeting and conference documents; 50,000 new acquisitions per year)
- usage statistics:
  - 270,000 searches per month
  - 25,000 distinct users per month (70% outside CERN)

Invenio at CERN: users

Use of personalization, CDS, August 2008:

- 12,082 registered users (39
- 2,230 email notification alerts
  - ... set up by 1,486 users (20
  - ... some alerts going to large user groups (ATLAS)
- 4,237 personal and group baskets
  - . . . set up by 3,168 users (36
  - . . . 528 shareable baskets, various access rights
  - . . . 50,871 records in baskets
  - . . . about 10 basket record additions per day

Invenio in the world

- About 25 advertized sites in production
- About 2,500,000 records worldwide altogether
- Interest from large document repositories and library networks
- examples:
  - Cini Foundation project, 1M digitized opera librettos and ancient works of art
  - NASA ADS, more than 7M astrophysics and related fields documents
  - EU Commission recent interest (new projects)
- Remark: CERN has no marketing office, but the technology transfer department is currently supporting a startup which goal it to provide services on top of InvenioIndico software

Some installations

- CERN Document Server (CERN, Geneva, Switzerland)
- International Linear Collider DOC (http://ilcdoc.linearcollider.org/)
- MeIND (HBZ NRW Koln Germany),
- INFOSCIENCE (EPFL, Lausanne, Switzerland)
- Aristotle University of Thessaloniki (Thessaloniki, Greece)
- RERO DOC (Martigny, Switzerland), EDUDOC (Swiss Education Server)
- PADIS (Universit La Sapienza, Rome, Italy)
- CAB UNIME (University of Messina, Italy)
- FYNU UCL (Universit catholique de Louvain, Belgium)
- McCamm on Group (UCSD, San Diego, USA)
- University of Applied Sciences of Fribourg (Fribourg, Switzerland)
- DDD (Universitat Autnoma de Barcelona, Spain)
- and many more...

Components

- Main programming language: Python (and C and Lisp)

Components

- Main programming language: Python (and C and Lisp)
- Runs on Apache using the Python module mod python, now replaced by WSGI

Components

- Main programming language: Python (and C and Lisp)
- Runs on Apache using the Python module mod python, now replaced by WSGI
- Uses MySQL RDBMS: take advantage of fully featured query language

Components

- Main programming language: Python (and C and Lisp)
- Runs on Apache using the Python module mod python, now replaced by WSGI
- Uses MySQL RDBMS: take advantage of fully featured query language
- Based on open standards (MARCXML, MARC21, OAIPMH, OpenURL...)

Components

- Main programming language: Python (and C and Lisp)
- Runs on Apache using the Python module mod python, now replaced by WSGI
- Uses MySQL RDBMS: take advantage of fully featured query language
- Based on open standards (MARCXML, MARC21, OAIPMH, OpenURL...)
- Medium to big data repositories

Components

- Main programming language: Python (and C and Lisp)
- Runs on Apache using the Python module mod python, now replaced by WSGI
- Uses MySQL RDBMS: take advantage of fully featured query language
- Based on open standards (MARCXML, MARC21, OAIPMH, OpenURL...)
- Medium to big data repositories
- Flexible in every layer: modular architecture

Components

- Main programming language: Python (and C and Lisp)
- Runs on Apache using the Python module mod python, now replaced by WSGI
- Uses MySQL RDBMS: take advantage of fully featured query language
- Based on open standards (MARCXML, MARC21, OAIPMH, OpenURL...)
- Medium to big data repositories
- Flexible in every layer: modular architecture
- Invenio home made Indexes

Index space design

- 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 $\rightarrowtail$ [rec1, rec2, . . .]
  - reverse : rec1 $\rightarrowtail$ [term1, term2, . . .]

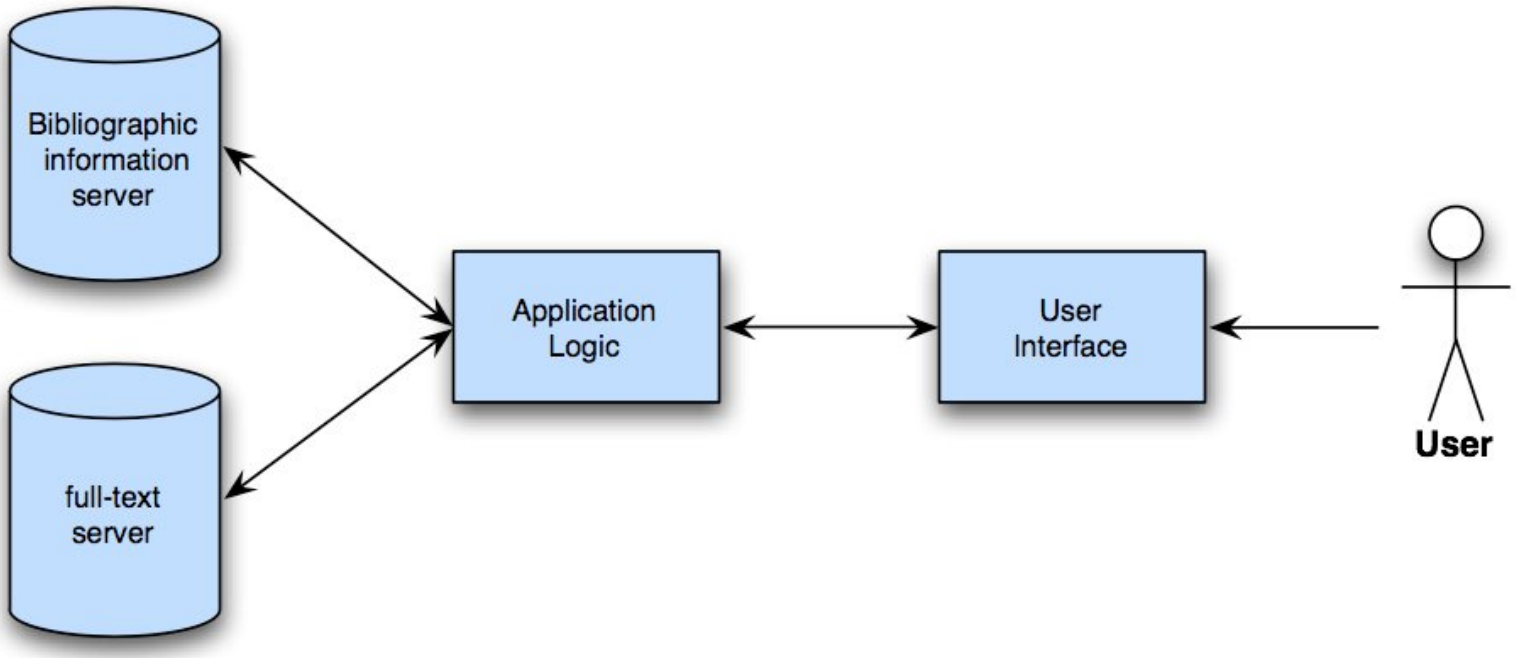
Performance stats

- Dual Xeon(HT) 3.06 GHz, SCSI Ultra320 with 1,000,000+ records, 550+ collections
- Indexing: total index size 11 GB, indexing time about 2 hours
  - 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

query	no. hits	search time
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

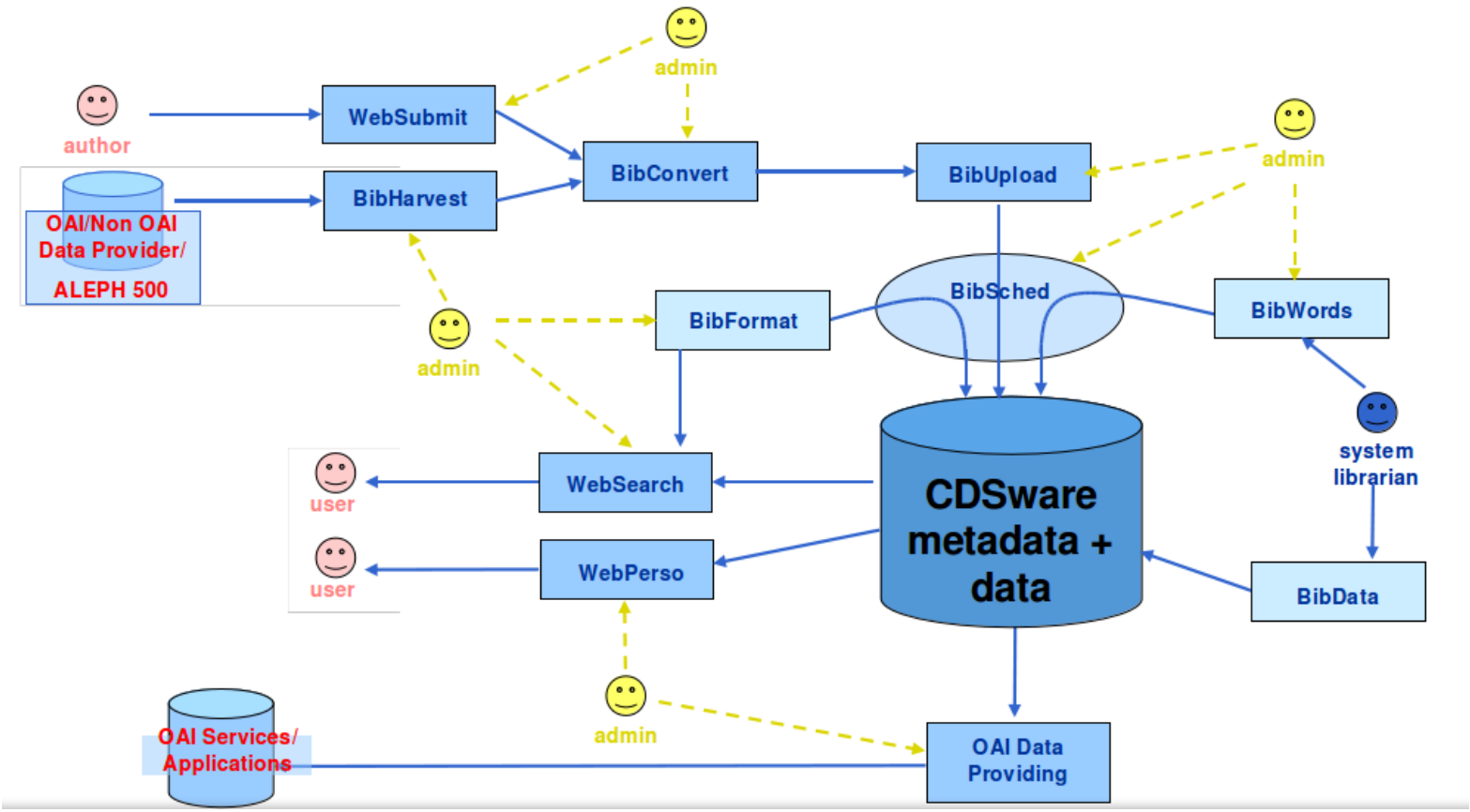
Simple view

- Application and DB Servers

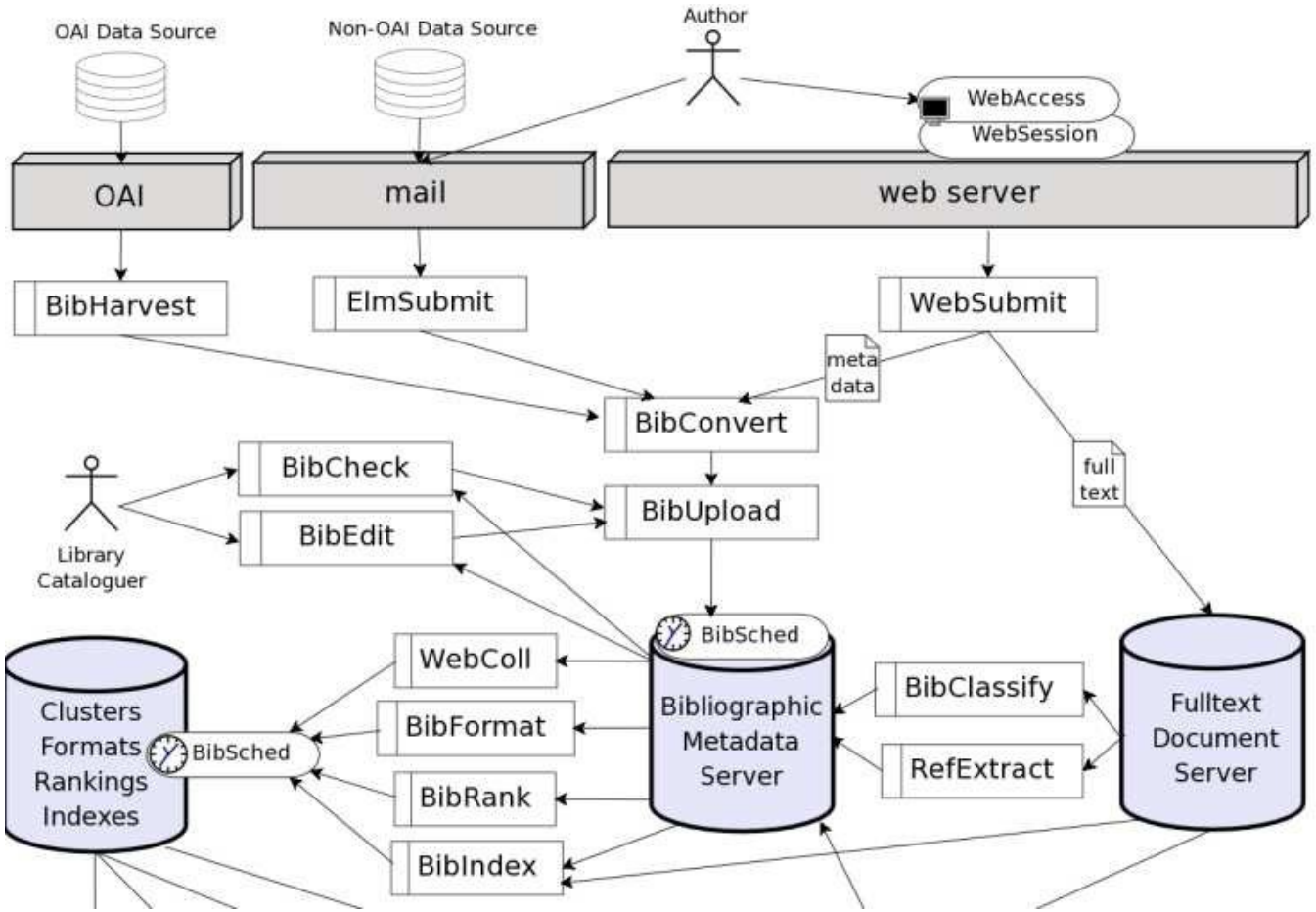


- Main load on the application server

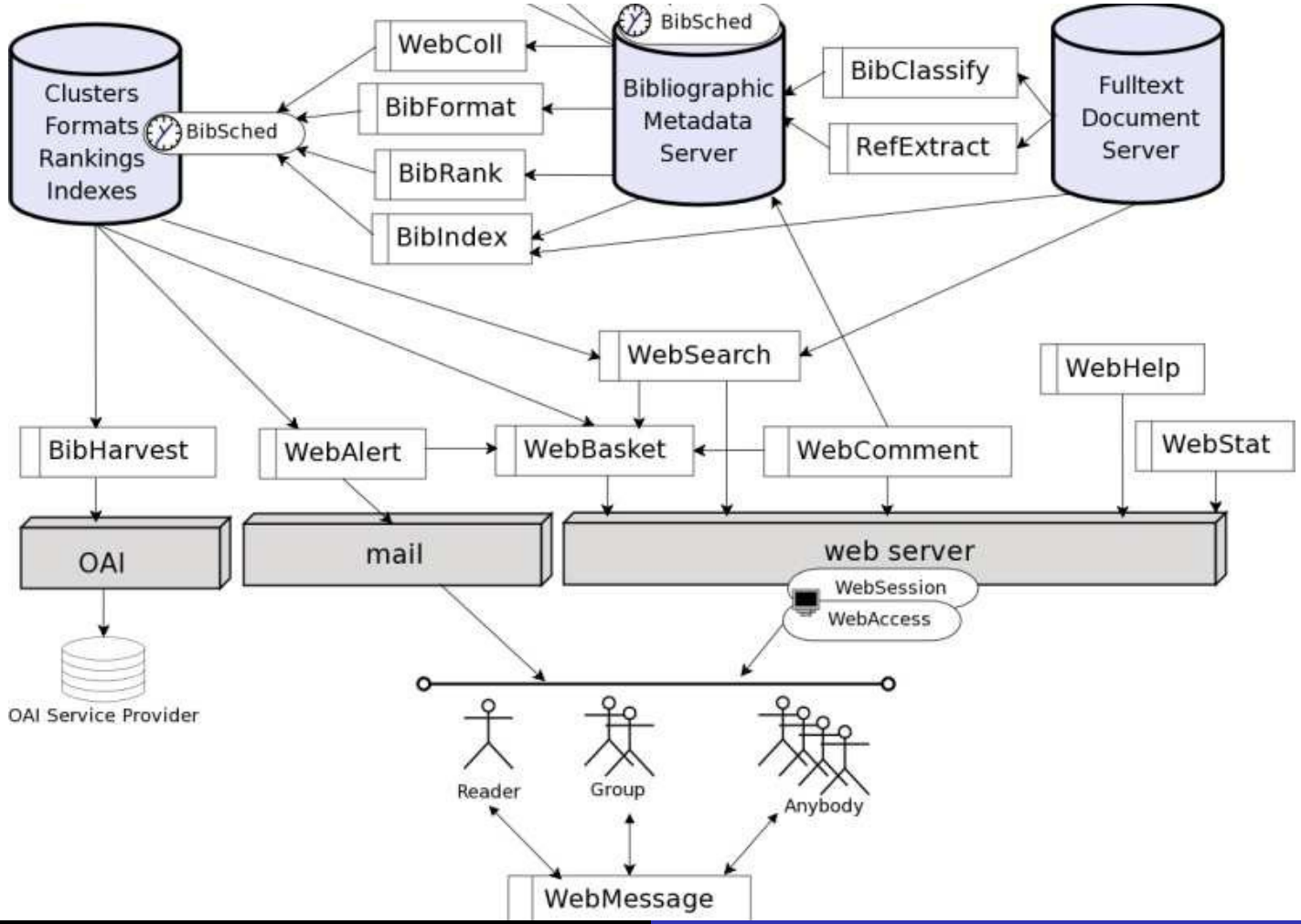
Simplified workflow



Module Overview I



Module Overview II



- CDS Invenio: a powerful, flexible solution suitable for the management of any collections of full text documents
- The driving force: an integrated vision for users
- Non Conventional Literature has required dedicated Software in the last 15 years: in-house developed system with natural growth development
- Today: the IR Software with the largest number of records world-wide
- Aims to enrich user experience by combining the best of the traditional library world with modern information retrieval technology
- Wide range of features: for users, librarians and administrators...

Links

- CDS Invenio consortium: http://cdsware.cern.ch
- Invenio at CERN: http://www.cern.ch
- Invenio for INSPIRE: http://inspire.cern.ch
- Contact: cds.support@cern.ch

Week organisation overview

- Tuesday: Invenio for Users
- Wednesday: Invenio for Lirarians
- Thursday: Invenio for Administrators
- Friday: Catchall session