CERN Document Server

Document Management System for Grey Literature in Networked Environment

Martin Vesely
CERN
Geneva, Switzerland

GL5, December 4 - 5, 2003
Amsterdam, The Netherlands

CERN Document Server

Overview

- Searching Scholarly Publications
  - Why not to use Google?
- Institutional Repositories
  - A natural way of document management at a place of the document origin
- Open Archives initiative (OAi)
  - develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content
  - enhances access to e-print archives as a means of increasing the availability of scholarly communication
- Protocol for Metadata Harvesting (PMH)
  - application-independent interoperability framework
- CERN Document Server
  - Implementation of an *institutional repository* and *information services* with *searching* and *harvesting* capabilities

Searching Scholarly Publications

"Electronic capabilities should be used to provide wide access to scholarship, encourage interdisciplinary research, and enhance interoperability and searchability.

Development of common standards will be particularly important in the electronic environment"

Principles for Emerging Systems of Scholarly Publishing

Tempe, Arizona, March 2-4, 2000

Z39.50

MARCXML MARC 21 XML Schema Official Web Site

S·F·X CONTEXT-SENSITIVE REFERENCE LINKING

CERN Document Server

Institutional Repositories

"Digital collections capturing, preserving and disseminating the intellectual output of a single or multi-university community"

SPARC
The Scholarly Publishing & Academic Resource Coalition
http://www.arl.org/sparc/

CERN Document Server

Open Archives Initiative

- Milestones of OAi:
  - Oct 1999, Santa Fe Convention
  - Nov 2000, OAi TC meeting at CERN
  - Jun 2002, OAi-PMH v.2.0 released

- Next:
  - CERN 3rd Workshop on Innovations in Scholarly Communication: Implementing the benefits of OAi
    12-14th February 2004 CERN, Geneva, Switzerland
    http://info.web.cern.ch/info/OAIP/

Protocol for Metadata Harvesting

- Services
  - Across institutional repositories

Institutional Repositories

- Application
  - e.g. search engine

- Metadata harvesting: OAi XML
  - Transfer: HTTP
    - other options
    - (+) HTTP widely deployed

Information Services

- Transport: communication subsystem
  - TCP/IP (internet)

Protocol for Metadata Harvesting

- Unified
  - XML Schema (structure)
  - HTTP transfer
  - Data encoding
  - Data flow control
  - Common transfer metadata format
- Independent
  - Storage technology
  - Local metadata format
  - Communication subsystem

CERN Document Server

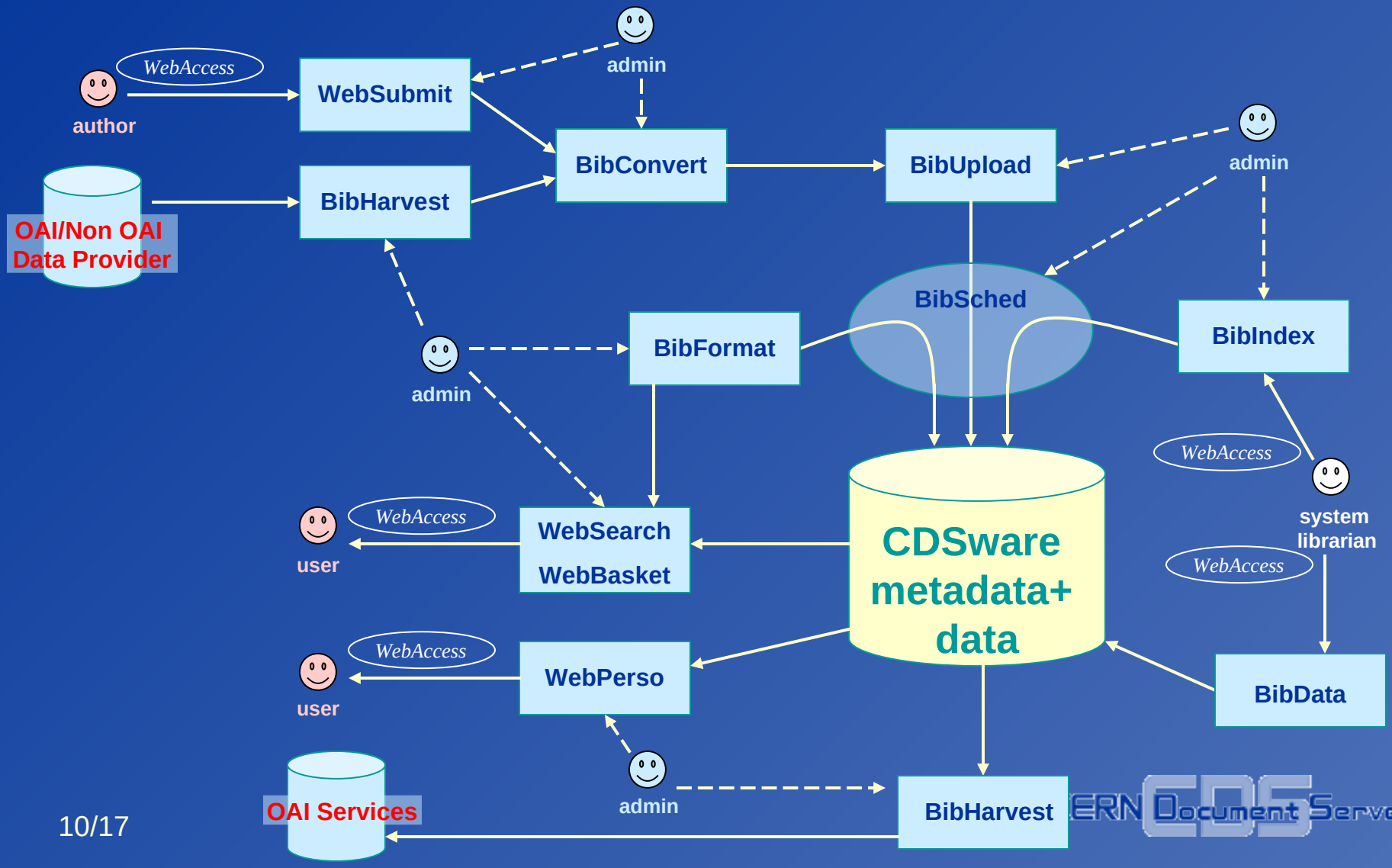
- CDS – digital library for HEP community
- CDSware in-house developed system
  - MySQL RDBMS, Apache, Python, PHP
  - MARC21 metadata format http://www.loc.gov/
  - Document submission (with flow control)
  - Multilingual: UNICODE

- CDSware is available as GPL http://cdsware.cern.ch/
- CVS repository access
- Free download and usage

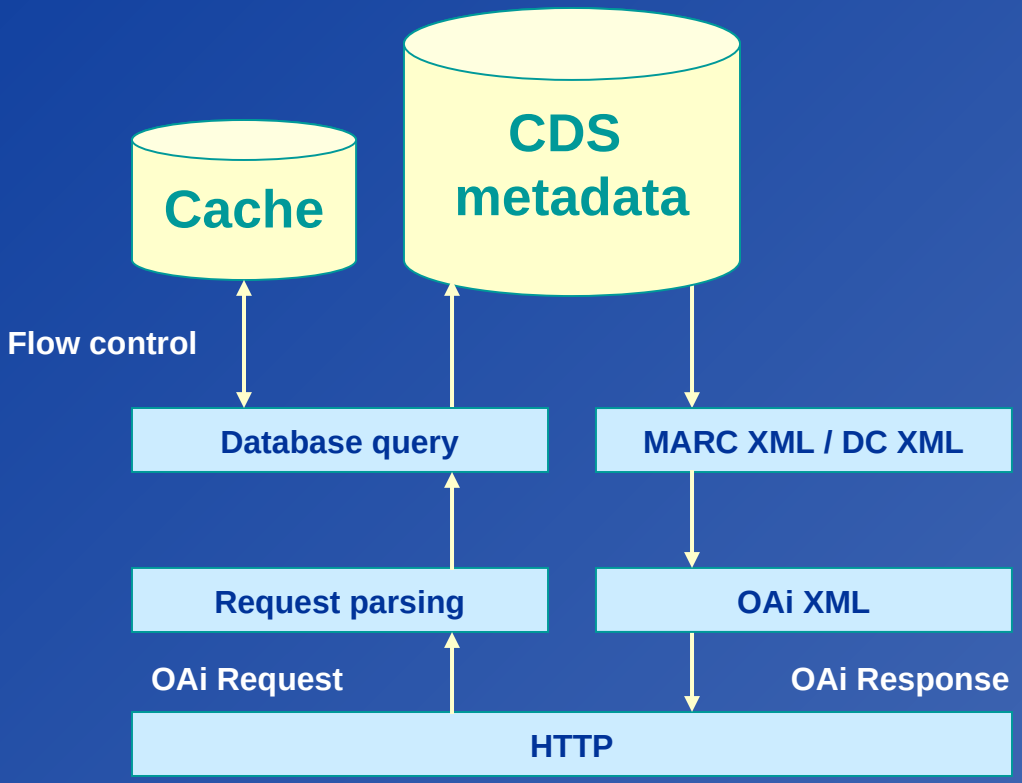
CDSware Search Engine

- Metadata organized into navigable collections
- In-house indexing technique to provide fast user-seen search times (fraction of a second for a typical query on a database upto size of 10^6 records)
- User friendliness, Google-like guidance
- Personalization:
  - Alert engine
  - User baskets
- Combined metadata/reference/fulltext searching

CDSware overview



CDSware OAi compliancy



CDSware References

- CDSware used or being considered by:
  - University of Missouri-Columbia , USA
  - Fundao Oswaldo Cruz (Ministry of Health) Rio de Janeiro, Brasilia
  - ISDN-ENSSIB, France
  - Montreal International
  - Bologna University, Italy
  - ETH Zurich, Switzerland
  - EPF Lausanne, Switzerland
  - UN Population Fund, New York, USA
  - Instituto de investigacions Electricas, Mexico
  - Casalini Libri, Italy
  - HBZ-NRW, Germany
  - SDSC, USA
  - Aristotle University of Thessaloniki, Greece
  - RERO: Consortium de toutes les bibliotheques publiques de Suisse Romande, Switzerland

CERN Document Server

Over **550,000** bibliographic records, including **220,000** fulltext documents,
of interest to people working in particle physics and related areas.
Covers preprints, articles, books, journals, photographs, and much more.

Search 629,741 records for:

any field SEARCH Browse

search tips :: advanced search

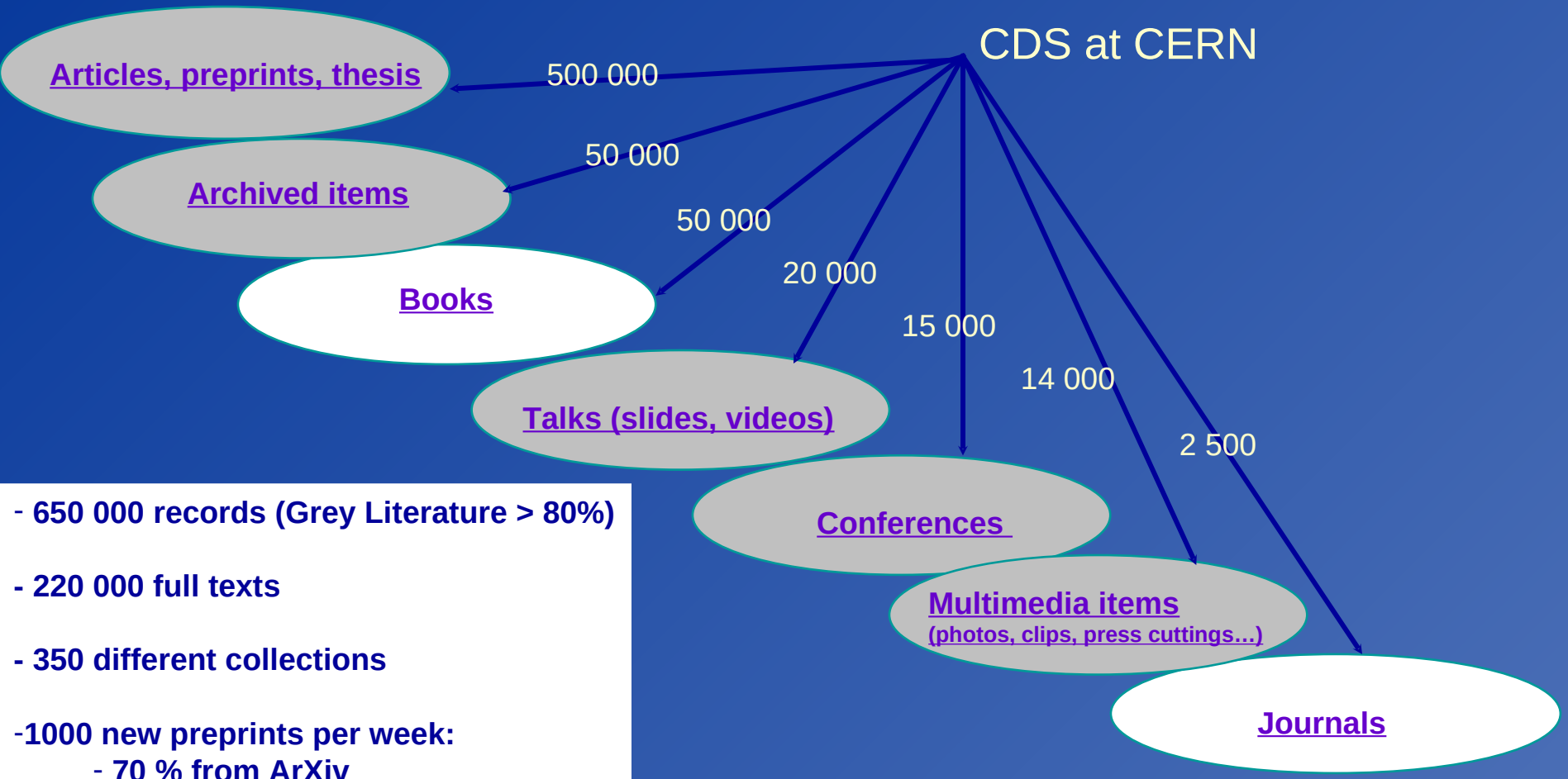
Narrow search:

- **Articles & Preprints** (504,827)
  Published Articles (154,216) Preprints (277,861) Theses (24,826) Reports (26,852) CERN Internal Notes (6,942) CERN Committee Documents (23,139)
- **Books & Proceedings** (50,185)
  Books (30,783) Proceedings (12,601) Standards (7,512)
- **Presentations & Talks** (13,477)
  Conference Announcements (11,963) Academic Training Lectures (508) Summer Student Lectures (192) General Talks (16) Videotapes (1,186)
- **Periodicals & Progress Reports** (2,986)
  Periodicals (2,342) Progress Reports (644)
- **Multimedia & Outreach** (14,666)
  Photos (4,881) Videos (105) Press Cuttings (5,350) Exhibition Objects (177) Posters (289) ATLAS eNews (84) Weekly Bulletin (1,783) HEP Institutes (917) Experiments at CERN (684) Internet Resources (396)
- **Archives** (47,950)
  CERN Archives (44,478) Pauli Archives (3,472) DSU Archives (701) SL Archives (1,026) AB Archives (163)

Focus on:

CERN Yellow Reports (1,059)
CERN Divisions (44,043)
Accelerator Sector (8,123) Administration Sector (21,688) Research Sector (11,863) Technology Sector (2,441)
CERN Experiments (7,819)
LEP Experiments (1,218) LHC Experiments (6,601)
CERN Projects (1,029)
LHC Project (1,029)
Associated Projects (792)
Geneva Research Collaboration (GRC) (792)

Documents at CERN



- 650 000 records (Grey Literature > 80%)
- 220 000 full texts
- 350 different collections
- 1000 new preprints per week:
  - 70 % from ArXiv
  - 5 % from CERN
  - 25 % from 80 other sources

CERN Document Server

Interoperability Issues

- Standardization efforts
  - XML Schemata and XSLT stylesheets have been specified (e.g. OAi-PMH)
  - Common metadata formats are defined (e.g. Dublin Core, MARC21)
- Semantic interoperability research
  - Structural approaches (e.g. RDF/XML)
  - Ontological Interoperability
  - Subject of research in DL

Conclusions

- Search engines for grey literature are being widely deployed and represent a central information service in scholarly communication
- Institutional repositories gain momentum and become dominant over disciplinary repositories
- Standardized frameworks for distributed and federated document processing have been established
- Information interoperability has been achieved on the syntactic and structural/schematic level, whereas semantic interoperability remains a research issue
- CDSware implementing OAi-PMH, freely available (GNU/GPL)

Contact

- CERN Document Server
  - http://cds.cern.ch/
  - http://cdsweb.cern.ch/
- CDSware sources and demo
  - http://cdsware.cern.ch/
  - http://cdsware.cern.ch:8000/DEMOPLUS/
- Contact
  - cds.support@cern.ch
  - martin.vesely@cern.ch