CERN Document Server Software

Martin Vesely
CERN
Geneva, Switzerland

Overview

- CERN Document Server Software
- Services within the CDS
- Providing CERN metadata
- OAI-PMH Implementation
- OAI-PMH Evaluation
- Conclusions

CDS Introduction

- CDS Software runs at CERN on:
  - 430.000 metadata records
  - 180.000 full text documents
  - 330 data collections
  - With ~15% CERN original documents
- Repository
  - MySQL database system
  - MARC21 format
  - Apache Web Server

OAI Sets

OAI repository

CDS Software is available under GPL

Services within the CDS

- Search engine
  - Google-like syntax
  - Designed for large data collections
  - Personal features (baskets, alerts)
- Document Submission (with flow control)
  - Peer reviewing for scientific notes
  - Approval of documents
  - 25 different types of submission
- Document Conversion Server
- Other services (Scan, Agenda, WebCast)

Data gathering before OAI

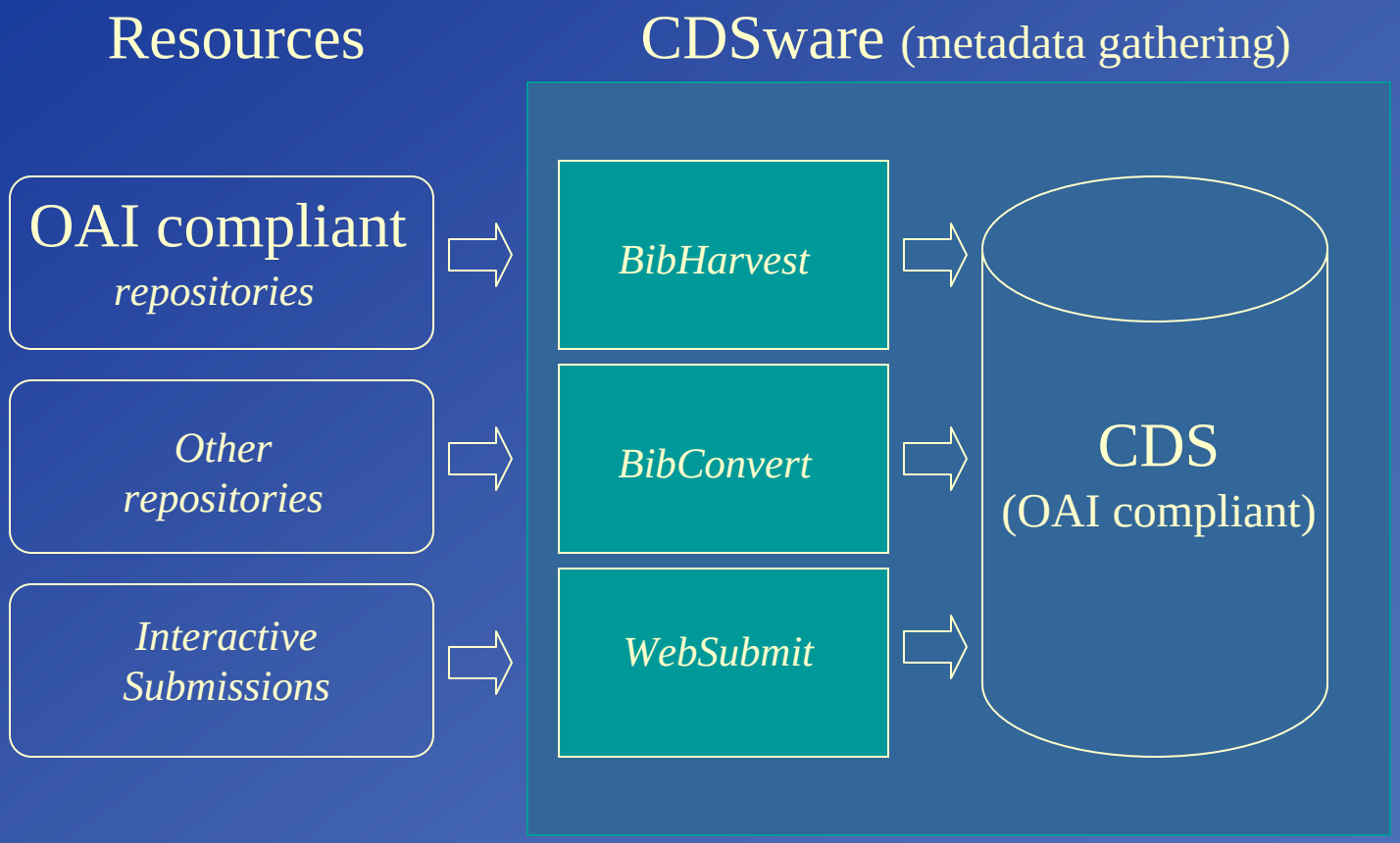
- Various types of resources
  - Structured metadata in various formats
  - Unstructured metadata (e.g. free text)
- Various transfer channels
  - http and ftp transfers, mail subscriptions
  - individual submissions
- Uploader application

XML Schema

XML

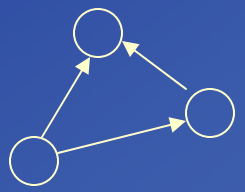
HTTP

Harvesting model…



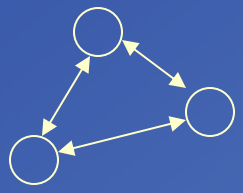
Providing CERN Metadata

- CERN as metadata repository
- Centralized vs. distributed model
  - Harvesting from multiple repositories
  - Two-way traffic / metadata sharing
- Hierarchical harvesting



Maintain
Most recent record

- Reciprocal harvesting



Identifiers of value added records

OAI-PMH Implementation

- CERN OAI Harvester (BibHarvest)
  - Modules
    - Metadata gatherer (crawler)
    - Scheduler
  - Python
- CERN OAI Repository (data provider)
  - Optional features
    - Data flow control
    - OAI Sets
  - Metadata Formats

Data flow control

- Resumption tokens (optional)
  - Expiration / lifetime
  - Transfer failure resistant (not guarantied)

Technique used	Notes
Complete snapshot (Cache all metadata fields)	+ database queried once - database replicated
Partial snapshot (no record caching)	+ saves resources + database queried once
Individual query (for each request)	+ saves resources - several database queries

OAI Sets

- Semantics
  - Defined by data provider
  - Description in XML container (opt. in v.2.0)

    human vs. machine readable
- Missing unification
  - Prevents cross-archive services
  - Sets by subject category

Metadata Formats

- Supported metadata formats
- Preferred metadata format
  - Information loss within metadata transfer
  - Conversion from native formats possible

DublinCore (only)	44 (64%)
RFC_1807	10 (14%)
MARC	8 (12%)
ETDMS	7 (10%)
OLAC	6 (9%)
Other (native)	9 (13%)
TOTAL	69

OAI-PMH Evaluation

- Advantages
  - Low-barrier access
  - Unified metadata transfer
  - Many optional features
  - “metadata brokering” support
- To be discussed
  - OAI identifiers
    - Persistent / dependent on enriched metadata
  - Application-level protocol proprietary solution
    - Direction of Web Services

Conclusions

- OAI-PMH v.2.0
- CDS Software is available under GPL
  - Implements both data provider and service provider
- Metadata transfer using pure oai_dc causes loss of information
- Cross-archive searches based on sets out of protocol scope

Further Information

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