CERN Document Server Software

1) Context
2) Interoperability
3) Submission
4) Search
5) Preservation

CERN, OAI3 Workshop, Geneva
Once upon a time …

CERN Library:

Mission of dissemination and long term keeping of HEP results

CONTEXT

THREE MAJOR CHANGES

1 - Computers start to be used in libraries

ALEPH Integrated system used at CERN

2 - First www server (Dec. 1990)

CDSware OAI-compliant distributed as GPL

1993: CERN preprint server → Library Server → CERN Doc Server

3 - Open Archive Initiative is launched (in 1999)
CDSware Architecture …

MySQL RDBMS
CDSware Indexes
Apache/Python
XML MARC
GNU GPL

Incremental organic-growth SW development model

CONTEXT
CDSware Interoperability

- **OAI Harvesting**
  - OAI Harvester: BibHarvest
  - Non-OAI Harvester: BibConvert
  - At CERN: more than 80 distinct sources are harvested

- **OAI Providing**
  - Records can be private, public and “OAI-public”
  - OAI Sets can be defined using any search criteria

- **Search Output Formats**
  - XML MARC; XML Dublin Core and more…
    - Any query is “OAI-ready”
    - Eg: OAI harvester could harvest only papers written by Ellis, J.
    - Eg: OAI harvester could harvest only title fields

- **Applications built on top of CDSware**
  - APIs to CDSware available

- **Connection with other Search Engines**
CDSware Submission process

- **Each collection can have its own submission policy**
  - Direct submission
  - Submission with monitoring
  - Submission with simple approval
  - Submission with peer review/refereeing and editorial board

- **Each collection can have its own record definition**
  - Metadata fields (mandatory, optional, controlled at input time…)
  - Full text formats
  - Revised versions

- **Each submission has its own process management**
  - With an HTML administration interface
  - To define submission screens
  - To define actions to be applied

- **‘Batch submission’ mode**
  - BibHarvest, BibConvert and BibUpload modules
CDSware Search

- **Google-like speed up to 1,000,000 records**
  - Web Application server ↔ DB server
  - DB insufficient: in-house performance-driven index design
  - Fast marshalling & fast set intersections:
    
    | 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 | 11,940  | 0.17 sec    |

- **Combined metadata/fulltext/reference search**
- **Multi-stage search guidance system**
- **Personalization: baskets, email alerts**
- **Navigable collection trees**
  - Primary and Virtual orthogonal views
- **Internationalization: multi-language interface**
CDSware: Long term preservation

- **CDSware at CERN**
  - “Certified Information System” (CIS)
  - Considered as a long term electronic archive
  - Hosts the official CERN Archives

- **MARC21 based: LOC standard**
  - XML MARC is the internal representation of CDSware records

- **Records deletion policy**
  - Record IDs never change

- **Full text automatically converted to PDF**
  - CERN Conversion server can be plugged in (GNU GPL)

- **Digital content disseminated… via OAI!**
CERN Document Server

Over 630,000 bibliographic records, including 250,000 fulltext documents, of interest to people working in particle physics and related areas. Covers preprints, articles, books, journals, photographs, and much more.

Search 650,252 records for:

- Articles & Preprints (521,214)
  - Published Articles (156,571)
  - Preprints (289,833)
  - Theses (26,534)
  - Reports (25,652)
  - CERN
  - Internal Notes (6,452)
  - CERN Committee Documents (24,105)

- Books & Proceedings (50,310)

Narrow search:

- 125,000 distinct hosts/clients in 2003
- 12,000 distinct hosts/clients per month
- 120,000 searches per month
- 5,000 OAI harvesting requests per month

- 650,000 different records
- 350,000 full texts
- 450 different collections
- 1000 new preprints per week:
  - 70% from ArXiv
  - 5% from CERN
  - 25% from 80 other sources

125,000 distinct hosts/clients in 2003

12,000 distinct hosts/clients per month

120,000 searches per month

5,000 OAI harvesting requests per month
CDSware: Conclusions

- Used in many places (dozen of installations)
- Dedicated support from CDS team (charged)
- Extending traditional library systems
- Designed to evolve
- Suitable for mid to large size repositories (1M recs)

http://cdsware.cern.ch