CERN Document Server:

An OAI-based solution for managing data collections

Jean-Yves Le Meur
CERN
Geneva, Switzerland

Starting Point

A physicist office

CERN-MI-9612016

CERN Contributions
to the open archive movement

- Hosting this workshop !
- Taking part into the technical committee
- Testing the versions of the protocol
- Delivering CERN documents via OAI

- And now: releasing CDSware as GPL
  CERN Document Server Software

CDSware at CERN covers:

- All particle Physics literature since 1950

and related areas documents: Astrophysics, Mathematics, Life at CERN…

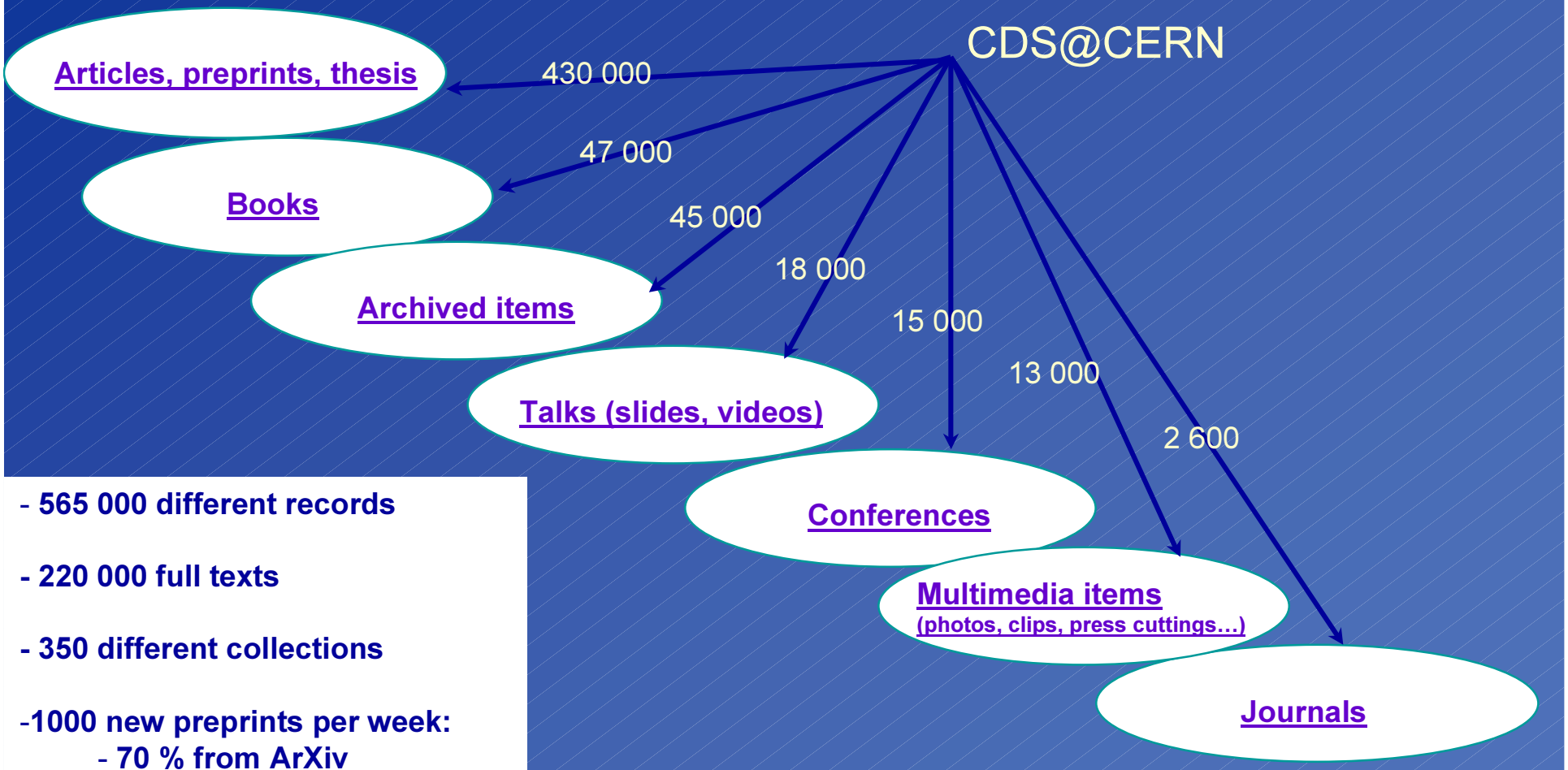
- 'Virtual' Collections:
  special views dedicated to an activity or a group.

  e.g: CERN Experiments collection (LHC, ATLAS, etc)
  CERN Divisions collections
  Customized views (Pauli collection)

- And It serves:

  156,000 distinct hosts/clients in 2001
  17,000 distinct hosts/clients per month
  1,000 "visits" and 3,500 searches per day
  50,000 "hits" and 1.5 GB net traffic per day

CDSware at CERN contains:



- 565 000 different records

- 220 000 full texts

- 350 different collections

-1000 new preprints per week:
  - 70 % from ArXiv
  - 5 % from CERN
  - 25 % from 80 other sources

http://cdsware.cern.ch

CDSware at CERN services:

CDSware general:

- First version released 1st of August 2002
  - All modules delivered as one single package
  - Distributed under GNU Public License.
  - Two mailing lists available, one for getting the news, and one for implementers discussions
  - Everything at http://cdsware.cern.ch

  80 000 lines of code !

- Built with:
  - MySQL, Apache, PHP, Python, WML
  - All customization & administration is web based

CDSware Featuring:

- WebSubmit: Submitting data
- BibHarvest: harvesting OAI repository
- BibConvert: harvesting non-OAI collections
- BibFormat: Formatting and linking records
- WebSearch: Searching metadata/citations/full text
- BibWord: Indexing metadata and full text
- WebAccess: Managing complex collection hierarchy
- WebPerso: Personalizing web access
- BibData: Modifying records (librarians only)

CDSware Direct Submit

- Web submission
  - by authors; by secretaries; by library staff
- Submission in steps and with control
  - Open; Monitoring; Approval; [Peer reviewing]
- Automatic Document conversion
- Automatic report number generation and stamping
- Multiple ‘post-submission’ functions. Eg:
  - Forward to distribution lists for advertising
  - Enable comments by peers
  - Modify submitted metadata
  - Send revised versions of full text
  - Extraction of citations
  - Extraction of author lists (when long)
  - [Extraction of keywords]

CDSware: harvesting strategy

- BibHarvest and BibConvert:
  allows to run massive importation of records
  - from OAI compliant data provider
  - from non OAI compliant provider
    - Template for describing the source to be uploaded
    - Template to describe the transformation of the source
- Always convert into OAI Marc XML, used as our internal record representation
- Also enable fetching full texts
- 95 % of CERN Library uploads !

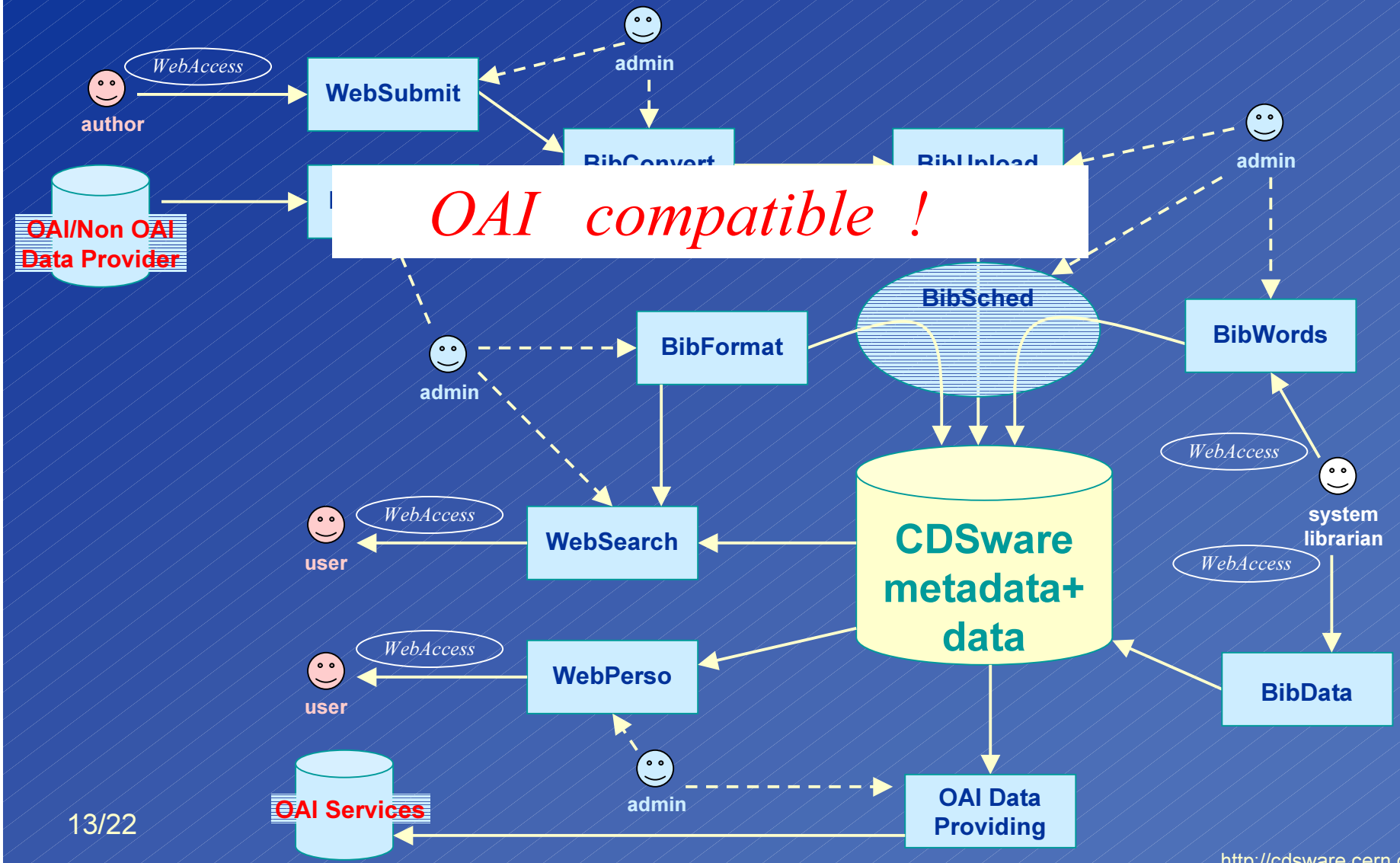
CDSware: linking strategy

- BibFormat: Flexible Formatting and Linking
  - All linking information separated from bibliographic information
  - Search Engine doesn't know anything about linking or formatting
  - Supports different types of link solving:
    - External linking → Just generate the link from stored rules
    - Internal linking → The link is always a file, it checks the existence, access, formats, etc
- First scenario:
  - Input: Bunch of records in OAI MARC XML
  - Output: Original XML record with its HTML version
- Second scenario:
  - Input: records in OAI MARC XML
  - Output: HTML version to be displayed or PHP to be saved to a file
- Egs: see http://doc.cern.ch/age?a02335
  - Links to full text
  - Links to articles or abstracts of e-journals

CDSware: Searching

- Google-like syntax and speed
- OAI functions implemented (v2.0)
- Marc21 representation database:
  - each field can be searched/browsed alone
- Full text, Citations and Metadata can be searched together with boolean operators
  - supported formats: PostScript, PDF, MS Word, MS Excel, MS PowerPoint
- Search options can be customized:
  - fields to be searched
  - sort options
  - formats of the records: html brief or detailed, xml oai dc+marc21, etc
  - splitting results by collections, with complex hierarchy
- Personalization options:
  - Baskets, alerts, layout

CDSware: Summary



OAI at CERN: our experience

The different points of view:

- Archivists
- Librarians
- Researchers
- Managers
- Computer scientists

OAI at CERN:
the archivist view

Do you really mean “Archive” ?…

- DC or MARC metadata is not enough:
  - OAIS (Reference Model for an Open Archival Information System).
- Important documents are printed.
  - Long term electronic preservation half-trusted
  - Need to run an “OA printshop” …

OAI at CERN:
the librarian view

Thank you but it does not solve everything !

- Look at a simple example:

  oai:arXiv:hep-th/0209017

OAI at CERN:
the librarian view - author exemple

- In subscription email:

  From: lukier@ift.uni.wroc.pl

  Author: J. Lukierski (Institute for Theoretical Physics, University of Wroclaw, Poland)

- With OAI GetRecord:

  <dc:creator>Lukierski, J.</dc:creator>

- In CERN Library:

  -email: <datafield tag="856" ind1="0" ind2=""> <subfield code="f">lukier@ift.uni.wroc.pl</subfield> </datafield>

  -author: <datafield tag="100" ind1="" ind2=""> <subfield code="a">Lukierski, J</subfield> </datafield>

  -affiliation: <datafield tag="909" ind1="C" ind2="1"> <subfield code="u"> Institute for Theoretical Physics, University of Wroclaw, Poland </subfield> </datafield>

OAI at CERN:
the librarian view - “comment” exemple

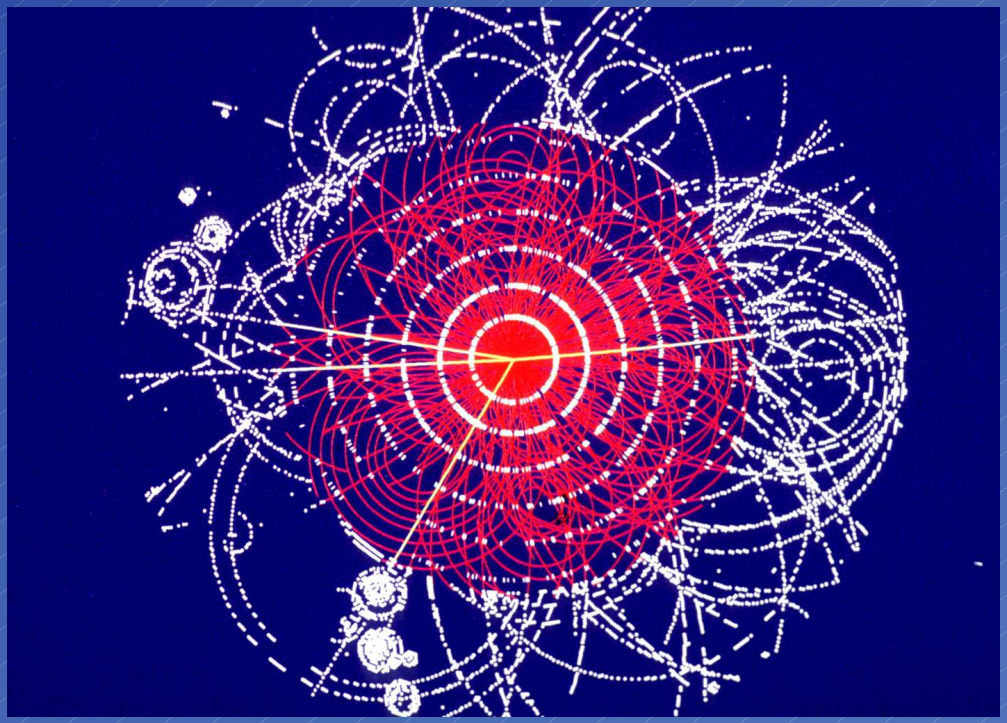
- With email or OAI GetRecord:
  - <dc:description>Comment: LaTeX, 9 pages, Invited talk at 11-th International Colloqium "Quantum Groups and Integrable Systems", June 2002, Prague, presented by J. Lukierski; in press in Proceedings Volume of Czech. J. Phys. vol. 52, (2002)</dc:description>

- In CERN Library:
  - Page number: <datafield tag="300" ind1="" ind2=""> <subfield code="a">9 p</subfield> </datafield>
  - Conference code: <datafield tag="909" ind1="C" ind2="K"> <subfield code="b">2314356</subfield> <subfield code="n">prague20020620</subfield> </datafield>

→ Appears in 11th International Colloquium on Quantum Groups and Integrable Systems , Prague, Czech Republic , 20 - 22 Jun 2002 (list conference papers)

OAI at CERN:
the researcher view

Where the hell is the Higgs Boson ?



CERN-DI-9506025

OAI at CERN:
the manager view

Does OAI make savings ?

- Some hope !
  - If one day it allows full high quality document harvesting → less maintenance
  - If one day it allows journal subscription cancellation
  - If one day it becomes a long term archiving solution
  - …
- But today ?
  - Let's get research grants (NSF, EC…) !

OAI at CERN:
the computer scientist view

- OAI: what a nice recipe !
  - Easy to cook
  - And still a lot to play with !
- A large community of OAI-adduct is born

Conclusion

- CERN will continue to be involved in the Open Archive movement by:
  - Providing, supporting, enhancing CDSware
  - Joining initiatives to promote the idea
- And let's hope it will be as successful as the open source movement…

Thank you.

Contact

- CERN Document Server
  - http://cds.cern.ch/
- CDSware sources, mailing lists, demo
  - http://cdsware.cern.ch/
- Contact
  - cds.support@cern.ch