CERN Document Server: Validation & OAI

WORKSHOP on the Open Archives initiative and Peer Review journals in Europe

Geneva, Switzerland

22 mars 2001

Document Server Background:

It contains:

- HEP documents: preprints, books, journals, photos, notes, presentations, meeting agendas, etc (25 types)
- 430 000 bibliographic records; 170 000 full text documents
- Aleph 300 library system (ExLibris)
- Customized Web interface
- A separate MySQL database for ‘non library’ documents

Users and Access

CDS is consulted by:

- Physicists at CERN and all over the world
- Distinct hosts counted :
  - Total of **127 000** distinct hosts in 2000
  - In average, 20 000 distinct hosts per month

CDS is loaded with:

- ~ 4 000 e-prints/month

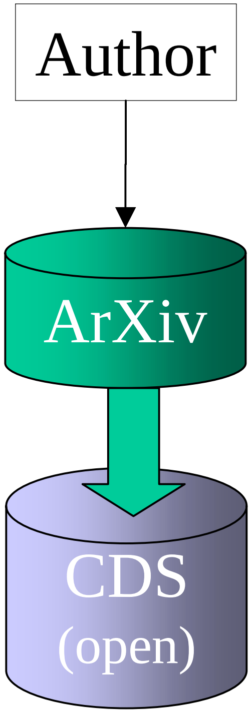
Metadata Acquisition @ CERN

- Manual (8%): collection of *scanned* documents
- Electronic:
  - Web & email *submission mechanism*
  - *Uploader* application for metadata transformation
- Long term storage system
- Five different “approval” approaches:

  from nothing to a complete review

1/ The Direct Way !

No Validation



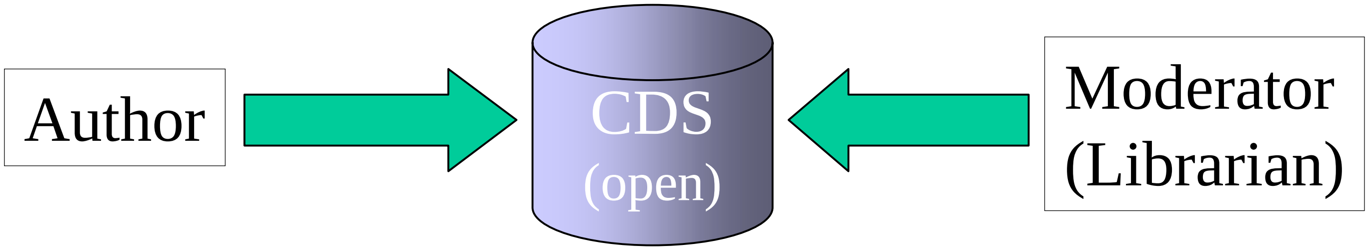
- ArXiv eprints

CERN author submits his paper to the ArXiv repository.

CDS gets it via the email subscription

2/ Moderation

- Open catalogue
- External submissions catalogue


Author

CDS
(open)

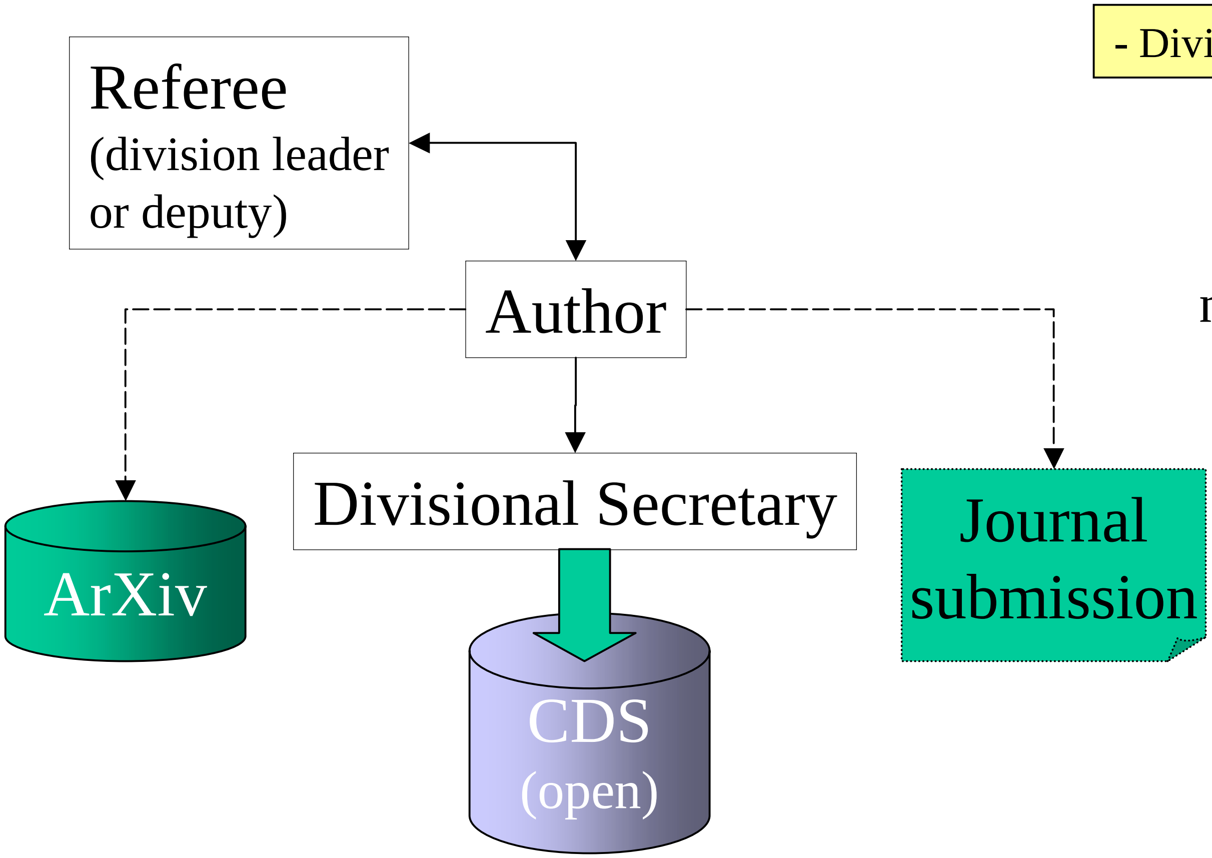
Moderator
(Librarian)

The author submits his paper
to CDS

A moderator decides whether
the report fits in the catalogue
or not

3/ Refereeing (manual)

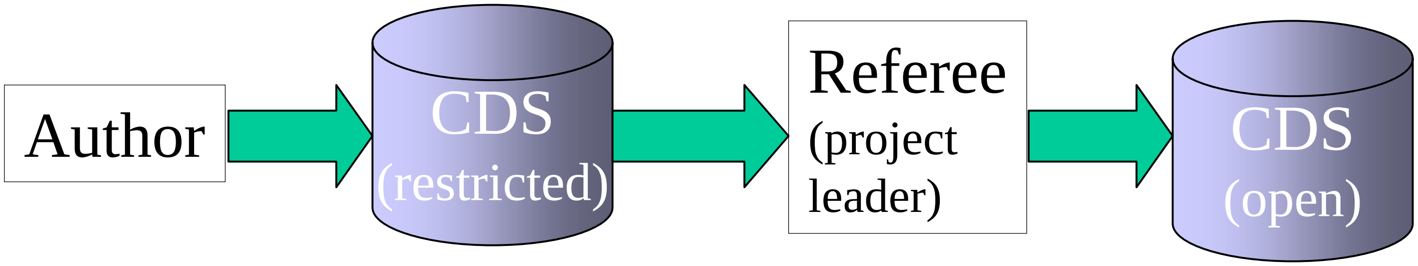
- Divisional Reports



The author gets an official CERN report number only if the referee validates his report.

4/ Refereeing (e-process)

- Collaboration Internal Notes
- Collaboration Pictures



Author

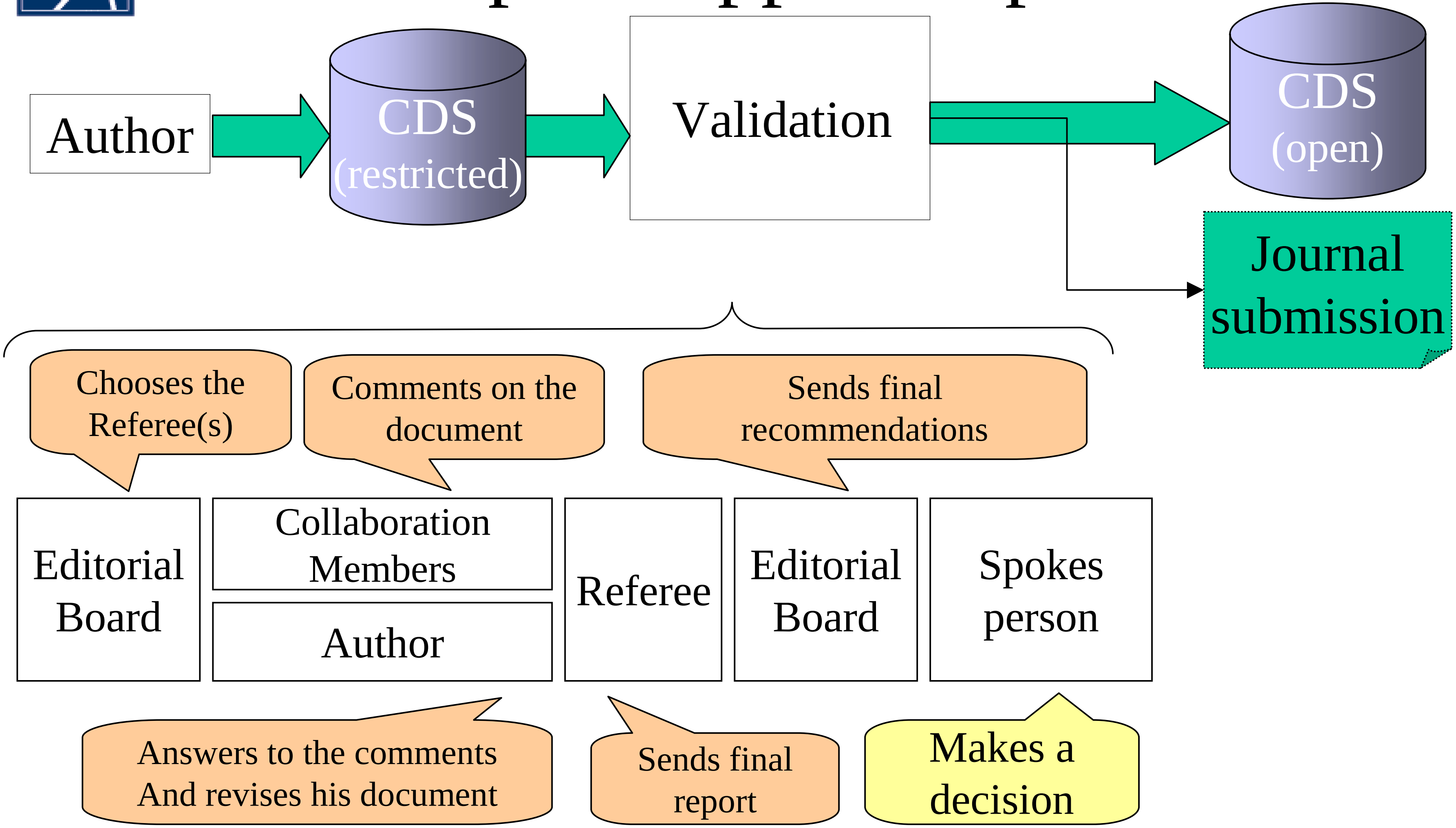
CDS (restricted)

Referee (project leader)

CDS (open)

The document is submitted electronically to CDS.
It is then kept in a restricted area as long as the referee does not approve it.

5/ Complete approval process



Validation and OAI

- CDS is ready for OAI compliancy as data provider
- In OAI philosophy: document quality is not recorded
- How to keep the value added by the validation?
- Simple solution: adding a quality label
  - Set-wide
  - Record-specific

Set-wide quality label

- Harvesting possible within OAI protocol
- Selective harvesting possible for service providers
- Problem #1: No qualitatively heterogeneous datasets -> proliferation of datasets
- Problem #2: Isolated record loses quality information

Record-specific quality label

- More flexible
- Keeps subject-driven sets
- Problem #1: needs cross-disciplinary standard quality label values
  - Solution: find a consensus
- Problem #2: selective harvesting of high quality documents impossible
  - Solutions: OpenURL, extended OAI protocol.

Conclusion

- Interest in quality labels:
  - – For data-providers:
    - availability of the validation information
  - – For service providers:
    - Possible harvesting of “high quality only” metadata
    - Relevance ranking according to quality labels

THE END

Can we afford to lose the validation information?

http://cds.cern.ch