CERN Digital Library

Strength, weakness and vision

Tim Smith
CERN
Geneva, Switzerland

Powered by:

INVENIO

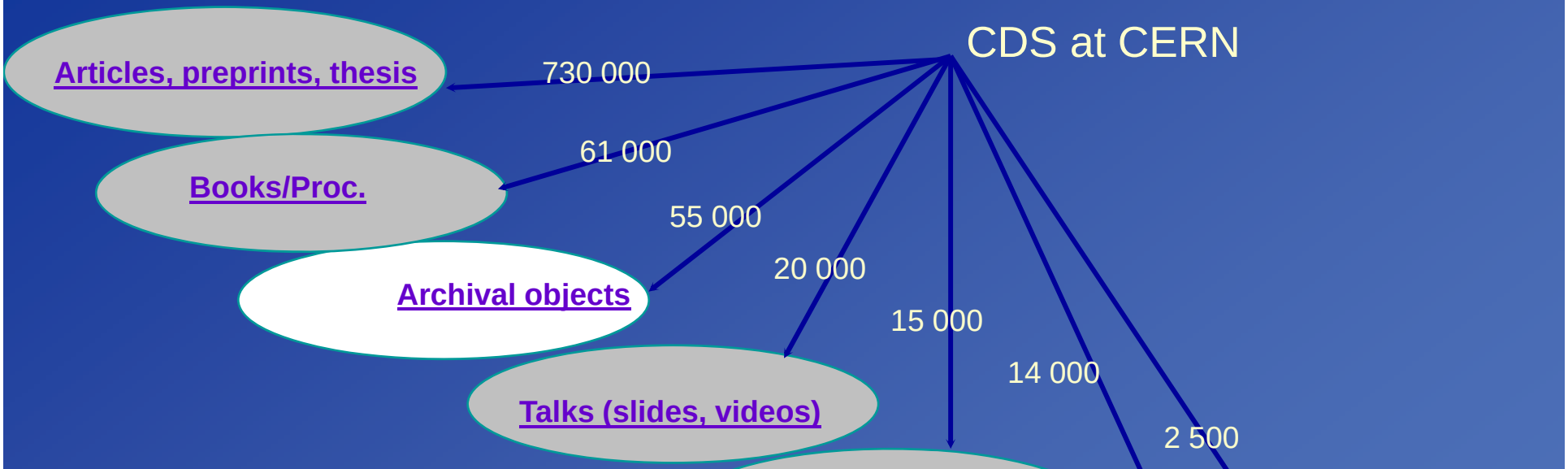
PPA Information Resource Summit,
SLAC, 10-12 May, 2007

An integrated vision for users

- Books from any library
- On site bookshop (including software)
  - Purchase of any book on behalf of any reader
- Standards
- Press cuttings
- Photos
- Exhibition objects
- Access to databases
- Access to electronic dictionaries and encyclopedias
- Access to eBooks
- Access to e-journals and back files

Most important: the key to the grey literature

Documents at CERN



- **900 000 records (Grey Literature > 80%)**

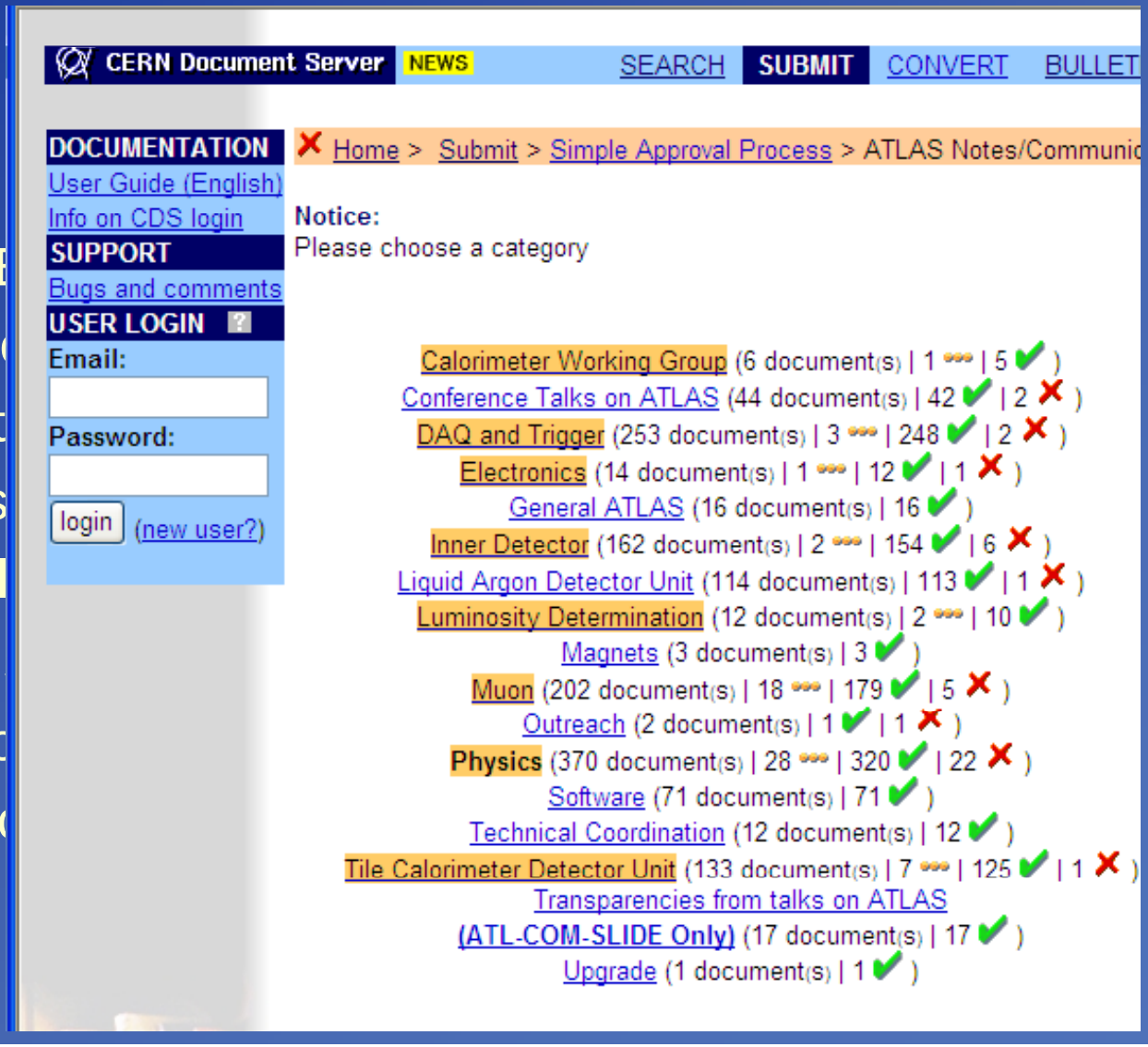
- **~50%available in fulltext**

-**A pertinent collection of CERN notes, thesis and multimedia objects**

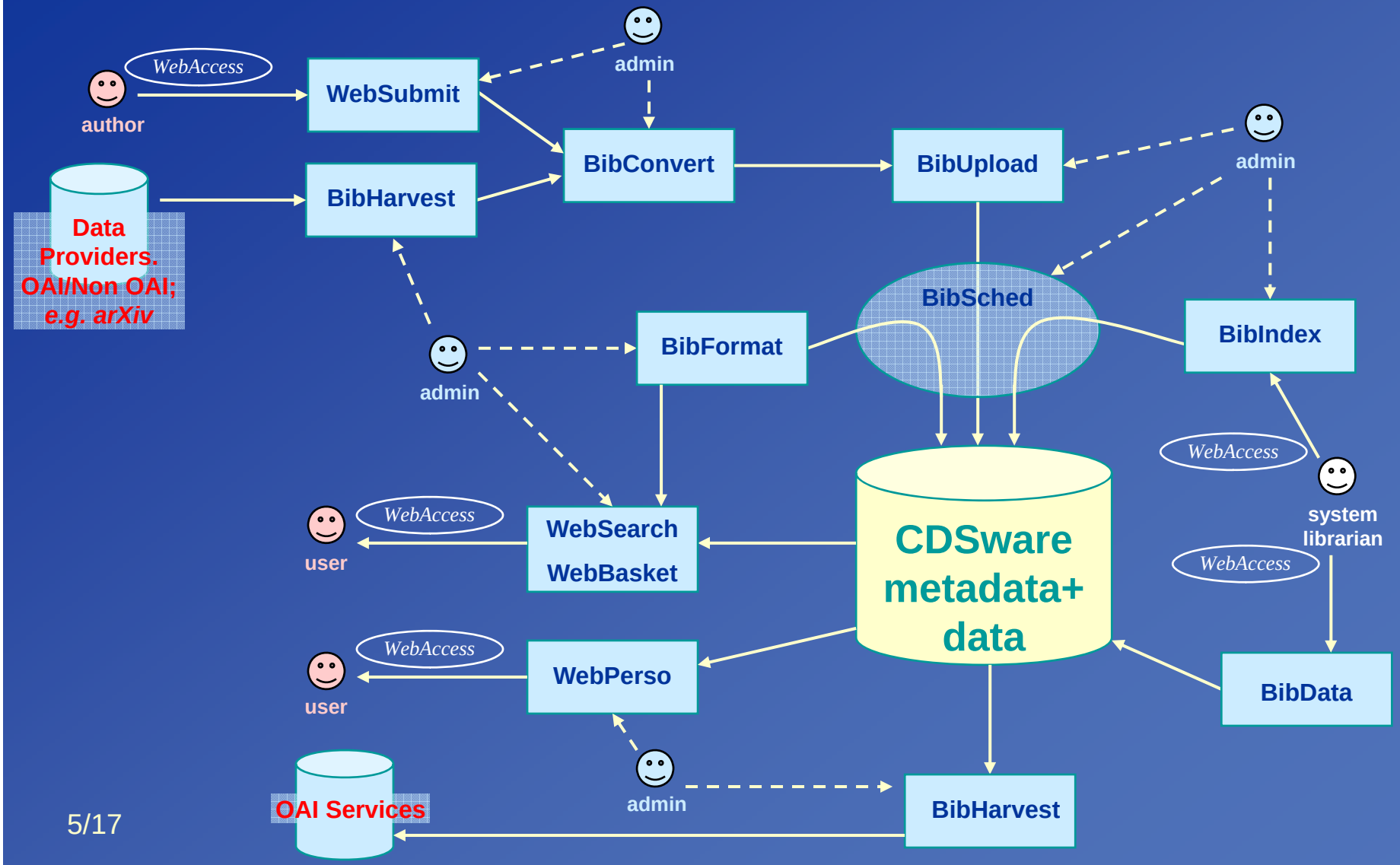
- **Source of input for article and preprints:**
  - **70 % from arXiv**
  - **5 % from CERN**
  - **25 % from a variety of sites**

CERN Document Server (CDS)

- CDS – digital
- CDS Invenio i
  - MySQL RDB
  - MARC21 m
  - Multi-format
  - Document s
  - Multi-lingual
- CDS Invenio i
  - Free downlo
  - Instances a



CDS Invenio overview



Key features

- Powerful search engine with Google-like syntax
  - includes parallel searching of external collections
  - Indico search – 30k conferences
- Combined metadata/reference/fulltext searching
  - For example, a query like *find all documents written by Ellis in 2002 that mention the term Higgs boson in the fulltext and that refer to Physical Review D 1997 papers* is very possible.
- Configurable portal-like interfaces for hosting collections
- User personalization
  - includes document baskets; email notification alerts; RSS feeds
- User collaboration, reviews, comments, knowledge sharing
- Multilingual interface available in 20 languages
- Automatic reference extraction; keywording
- Open Archive Initiative protocol compliant (metadata harvesting)

Strengths

- Cutting-edge technology
- Powerful search possibilities; *e.g.* combined searching in metadata and fulltext
- Long-time support; ongoing R&D for further improvement/development
- Scalable solution

Weaknesses

- The user interface should be better tuned with the needs of the community
- The coverage should be extended in time and depth in order to satisfy expressed needs from the user base
- Serves well as an institutional repository, however, cannot compete with SPIRES as the one-stop HEP information source due to lack of:
  - data quality, completeness
  - user perception, tradition etc.

Vision

- In collaboration with all stakeholders, build a complete HEP information platform, reducing current redundant efforts:
  - integrating the content of present HEP repositories and databases to host the entire body of HEP metadata and the full-text of all HEP OA publications, past and future;
  - developing and deploying new approaches to automatic subject detection, key-wording and classification of articles;
- Enabling new full-text and data-mining applications on all HEP publications:
  - detecting **relations** between documents carrying similar information, or produced by an overlapping set of authors, or referencing a similar set of papers;
  - building freely-accessible complete datasets to exercise new hybrid metrics to measure the impact of articles and authors and evaluating the scientific production of research groups;
  - identifying potential referees for submitted papers using automatic content analysis;
  - extracting numerical information from figures and tables within published articles.
- Demonstrate and deploy Web2.0 applications in the domain of sciences:
  - enabling the possibility to review and comment on articles, adding links to additional documents or other digital objects;
  - providing collaborative tools for effective management of co-authorship within distributed collaborations;
  - introducing community-based alternatives to the established peer-review system.

Concluding observations

- The different players in HEP information have maintained parallel systems already too long; we should join forces and work smarter in order to achieve better results with less effort
- Today users cannot easily profit from the strength of all the systems; more often users buy into one system and remain linked to its qualities and weaknesses throughout their career
- The PPA Information Resource Summit is a golden opportunity to make a first step toward a new era