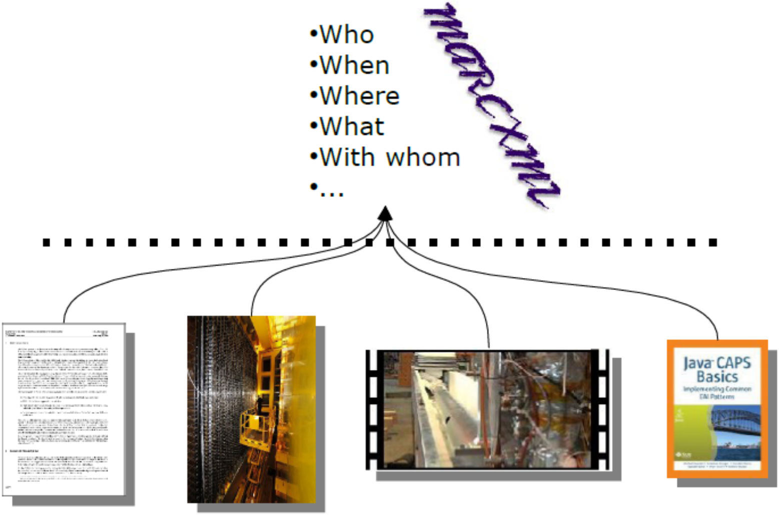
CDS-Invenio for librarians

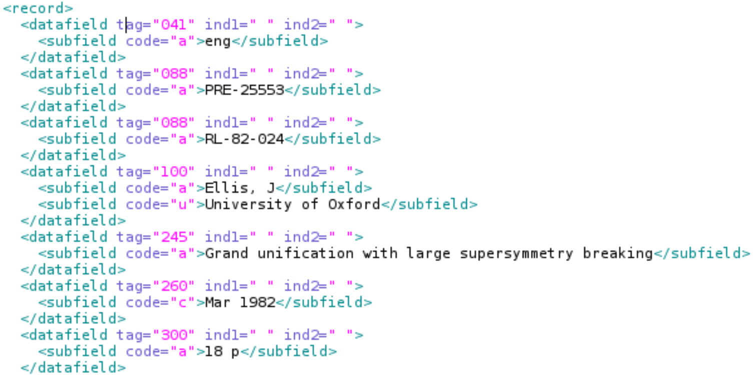
Jean-Yves Le Meur - CERN Document Server Project Leader

September 22, 2009

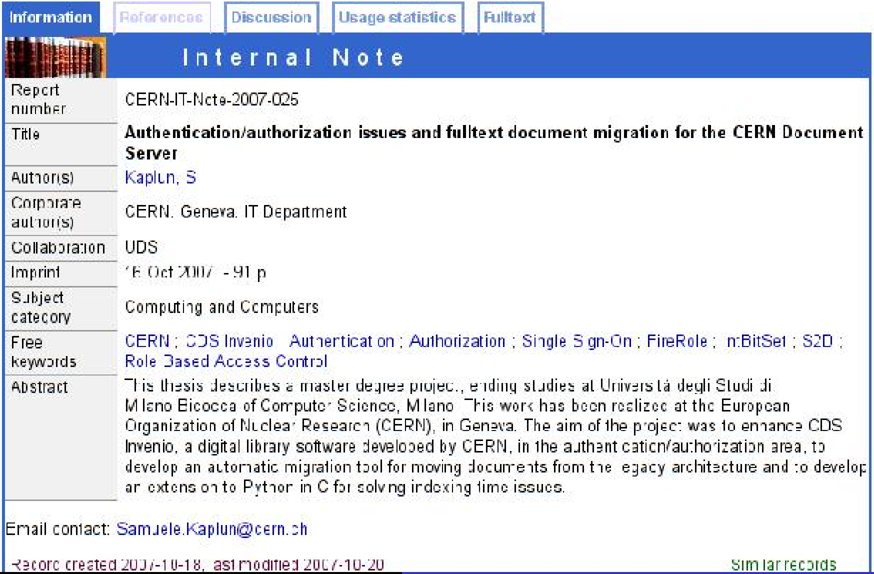
- Record definition
  - What is a record ?
  - Internal representation of records: MARCXML
  - The "life" of a record
- Document Submission by Authors
  - Web Submission
  - Submissions at CERN
  - Email Submission
- Batch Acquisition of records
  - Acquisition from other systems
  - Harvesting Open Access Compatible sites
  - Converting Records
  - Harvesting and Converting Fulltexts
- Editing tools for librarians
  - Cataloguing tools overview
  - Bibliographic Information Edition



- Who
- When
- Where
- What
- With whom
- ...

- MARC Standard exists since the 60s
- established by the Library of Congress, it is used by most library systems to describe bibliographic information of records
- very large number of fields, and freedom to "invent" new ones
- 

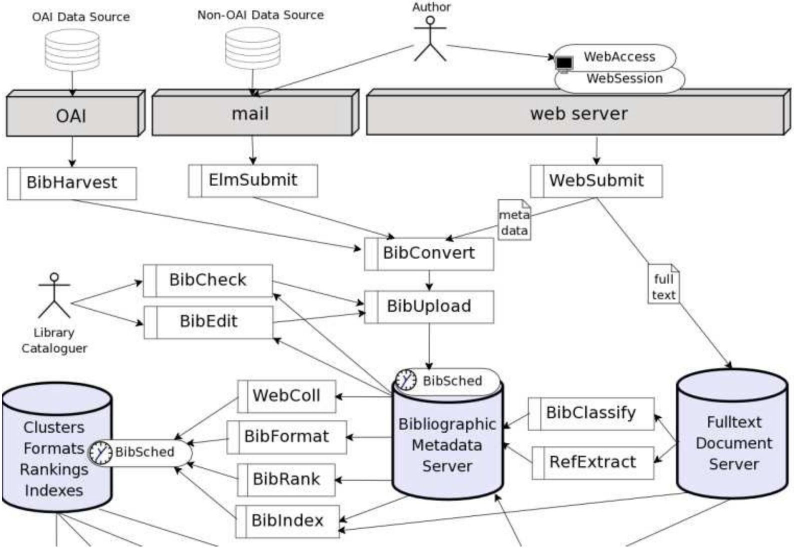
```
<record>
  <datafield tag="041" ind1=" " ind2=" ">
    <subfield code="a">eng</subfield>
  </datafield>
  <datafield tag="088" ind1=" " ind2=" ">
    <subfield code="a">PRE-25553</subfield>
  </datafield>
  <datafield tag="088" ind1=" " ind2=" ">
    <subfield code="a">RL-82-024</subfield>
  </datafield>
  <datafield tag="100" ind1=" " ind2=" ">
    <subfield code="a">Ellis, J</subfield>
    <subfield code="u">University of Oxford</subfield>
  </datafield>
  <datafield tag="245" ind1=" " ind2=" ">
    <subfield code="a">Grand unification with large supersymmetry breaking</subfield>
  </datafield>
  <datafield tag="260" ind1=" " ind2=" ">
    <subfield code="c">Mar 1982</subfield>
  </datafield>
  <datafield tag="300" ind1=" " ind2=" ">
    <subfield code="a">18 p</subfield>
  </datafield>
```



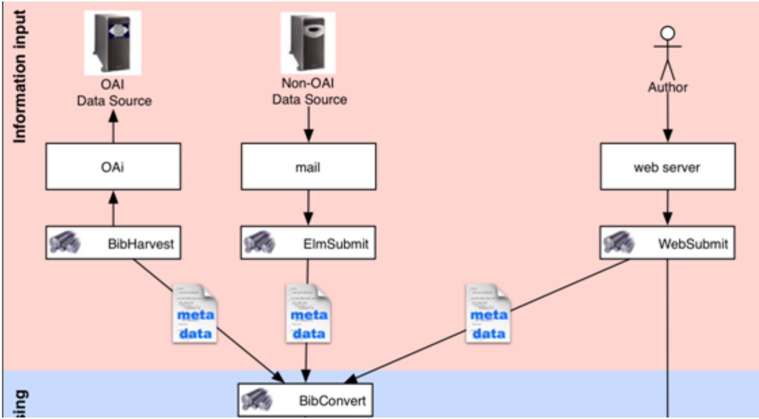
Report number	CERN-IT-Note-2007-025
Title	**Authentication/authorization issues and fulltext document migration for the CERN Document Server**
Author(s)	Kaplun, S
Corporate author(s)	CERN. Geneva. IT Department
Collaboration	UDS
Imprint	16 Oct 2007. - 91 p
Subject category	Computing and Computers
Free keywords	CERN ; CDS Invenio ; Authentication ; Authorization ; Single Sign-On ; FireRole ; intBitSet ; S2D ; Role Based Access Control
Abstract	This thesis describes a master degree project, ending studies at Università degli Studi di Milano Bicocca of Computer Science, Milano. This work has been realized at the European Organization of Nuclear Research (CERN), in Geneva. The aim of the project was to enhance CDS Invenio, a digital library software developed by CERN, in the authentication/authorization area, to develop an automatic migration tool for moving documents from the legacy architecture and to develop an extension to Python in C for solving indexing time issues.

Email contact: Samuele.Kaplun@cern.ch

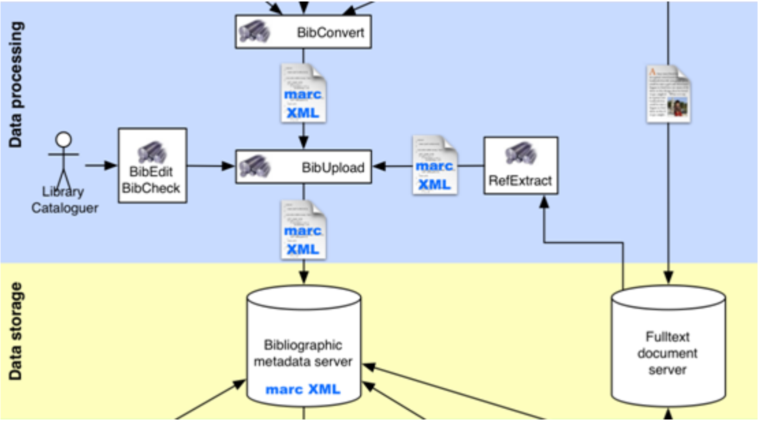
Record created 2007-10-18, last modified 2007-10-20 Similar records


OAI Data Source
Non-OAI Data Source
Author
WebAccess
WebSession
OAI
mail
web server
BibHarvest
ElmSubmit
WebSubmit
meta data
BibConvert
full text
Library Cataloguer
BibCheck
BibEdit
BibUpload
BibSched
WebColl
BibFormat
BibRank
BibIndex
Clusters
Formats
Rankings
Indexes
BibSched
Bibliographic Metadata Server
BibClassify
RefExtract
Fulltext Document Server

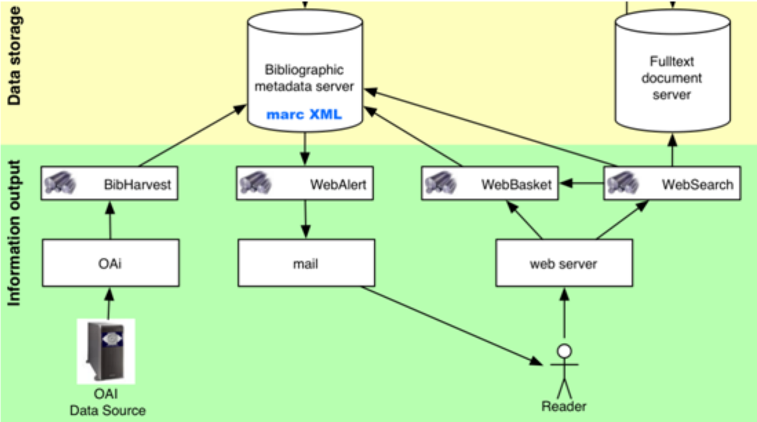
Information Input



Data Process and Storage



Information Output

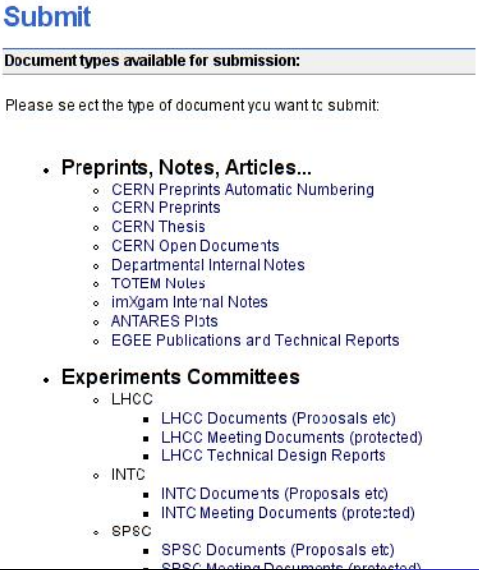


Web submissions (I)

- 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

Web submissions (II)

- Each submission has its own process management
  - It can be configured with an HTML administration interface
  - To define submission screens
  - To define actions to be applied when the document is transferred
  - Examples:
    - When finishing the submission of a videotape by video service, the label is created (PDF) to be stick on the tape
    - When a note is submitted by a collaboration, members of the collaboration are immediatly notified by email for comments



Submit

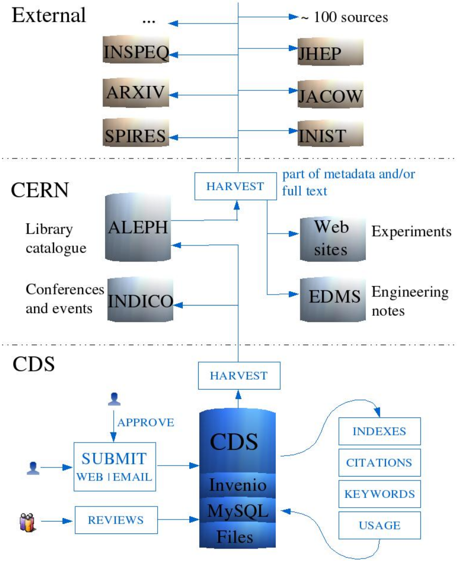
Document types available for submission:

Please se ect the type of document ycu want tc submit:

- **Preprints, Notes, Articles...**
  - CERN Preprints Automatic Numbering
  - CERN Preprints
  - CERN Thesis
  - CERN Open Documents
  - Departmental Internal Notes
  - TOTEM Notes
  - imXgam Internal Notes
  - ANTARES Plots
  - EGEE Publications and Technical Reports
- **Experiments Committees**
  - LHCC
    - LHCC Documents (Proposals etc)
    - LHCC Meeting Documents (protected)
    - LHCC Technical Design Reports
  - INTC
    - INTC Documents (Proposals etc)
    - INTC Meeting Documents (protected)
  - SPSC
    - SPSC Documents (Proposals etc)
    - SPSC Meeting Documents (protected)

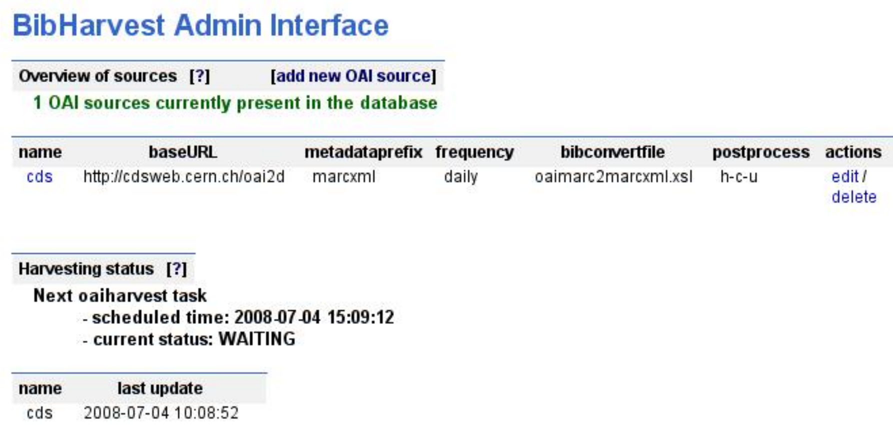
Email Submission

- Users can send their document as attachement by email
- The destination must be a dedicated Invenio email address (eg: submit@invenio.cern.ch)
- Subject and Body of the email contain the bibliographic information
- It must be well formatted to be accepted by the server
- The server returns aknowledgment of the succesful submission
- Almost not used at CERN, but it could become the most efficient system for end-users


External
...
~ 100 sources
INSPEQ
JHEP
ARXIV
JACOW
SPIRES
INIST
CERN
HARVEST
part of metadata and/or full text
Library catalogue
ALEPH
Web sites
Experiments
Conferences and events
INDICO
EDMS
Engineering notes
CDS
HARVEST
APPROVE
SUBMIT
WEB | EMAIL
REVIEWS
CDS
Invenio
MySQL
Files
INDEXES
CITATIONS
KEYWORDS
USAGE

Open Access Harvesting

- Collaborative framework with Data and Service Providers
- Information interoperability
- Value added processing (metadata brokering)
- OAI-PMH protocol: Data Provider $< -- >$ Service Provider
- Invenio has a Web Interface to manage Harvesting and Dissemination via OAI-PMH protocol



BibHarvest Admin Interface

Overview of sources [?] **[add new OAI source]**

1 OAI sources currently present in the database

name	baseURL	metadataprefix	frequency	bibconvertfile	postprocess	actions
cds	http://cdsweb.cern.ch/oai2d	marcxml	daily	oaimarc2marcxml.xsl	h-c-u	edit / delete

Harvesting status [?]

Next oaiharvest task
- scheduled time: 2008-07-04 15:09:12
- current status: WAITING

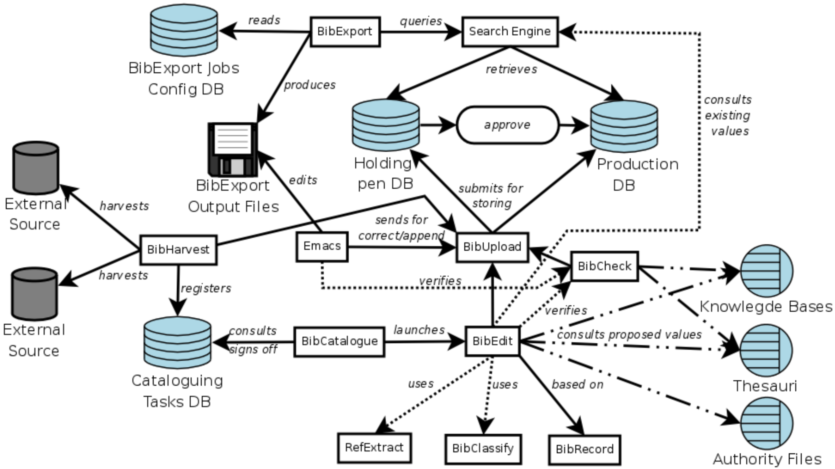
name	last update
cds	2008-07-04 10:08:52

Converting Records

- It can be performed using ad hoc scripts (e.g: using yaz), to deal with specific formats
- Lot of converters from *anything* to MARCXML already exist (from LoC site)
- Invenio proposes two internal ways of converting incoming records:
  - XSLT: most standard tool to transform XML files
  - BibConvert descriptive tagging
- BibConvert
  - You first describe your source file with tags
  - You then describe your target file with the same tags
  - The converter moves the information within the new scheme

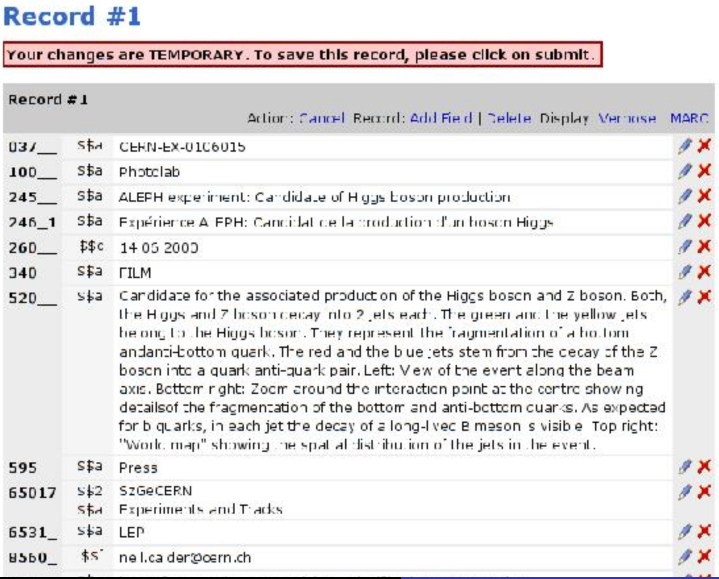
- The FFT Protocol: Fulltext File Transfer
  - Additional "FFT" tag in MARC records when submitted to BibUpload: $< datafieldtag = "FFT" ind1 = "" ind2 = "" >< subfieldcode = "a" > http : //cdsware.cern.ch/download/invenio - demo - site - files/0101431.ps.gz < /subfield >< /datafield >$
- Fulltext files can be uploaded at the same time as the metadata
- If files are scanned files (bitmap), OCR process can be started automatically
  - to recognize text and index it
  - to let users copy/paste text inside the files
- Files are managed with a dedicated module: BibDocFile

- goal: reproducing traditional library systems cataloguer-level functionality
- record editing interface
- record checking tools
- record maintenance tools
- record inputting workflow
- record harvesting workflow
- knowledge bases


reads
BibExport
queries
Search Engine
BibExport Jobs Config DB
produces
retrieves
consults existing values
approve
Holding pen DB
Production DB
BibExport Output Files
edits
External Source
harvests
submits for storing
sends for correct/append
BibHarvest
Emacs
BibUpload
BibCheck
Knowlegde Bases
harvests
verifies
registers
External Source
verifies
consults
signs off
BibCatalogue
launches
BibEdit
consults proposed values
Thesauri
Cataloguing Tasks DB
uses
uses
based on
RefExtract
BibClassify
BibRecord
Authority Files

Editing Tool for cataloguers

- Direct access to any record (authorized)
- Access control on who can edit what
- Web interface with Javascript (JQuery)
- Designed for distributed cataloguing
- Record locking mecanism to avoid parallel modifications



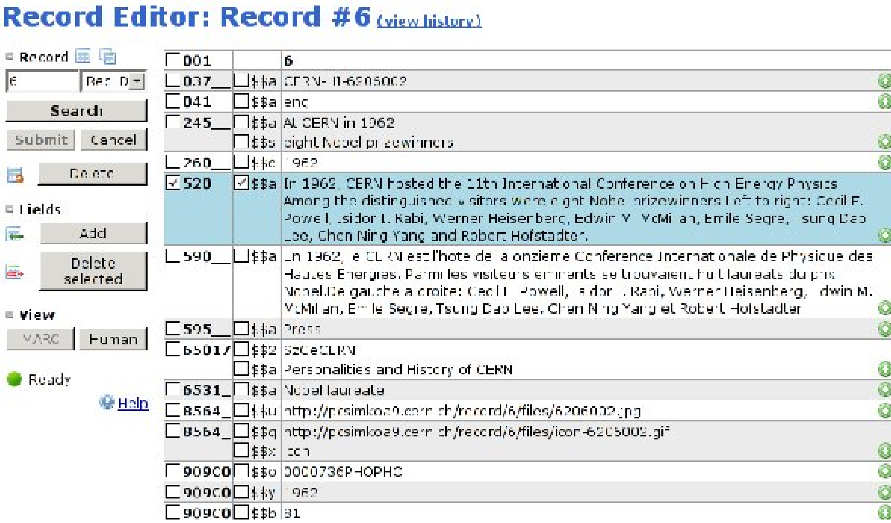
Record #1

Your changes are TEMPORARY. To save this record, please click on submit.

Record #1

Action: Cancel Record: Add Field | Delete Display: Verbose MARC

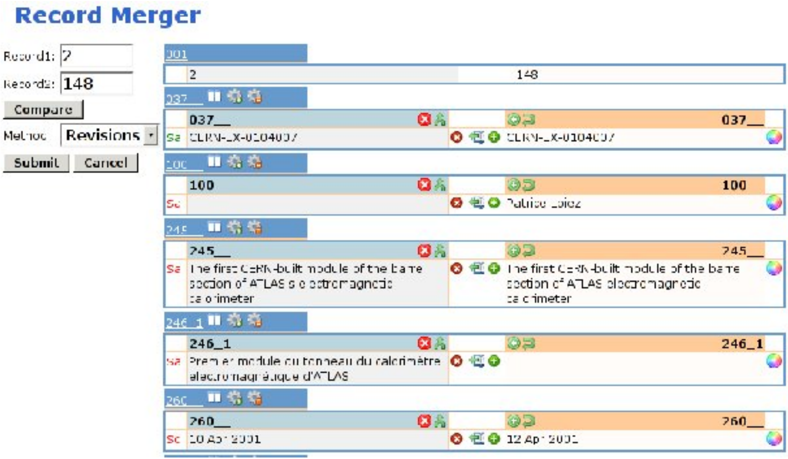
037__	$$a	CERN-EX-0106015	✎ ✗
100__	$$a	Photolab	✎ ✗
245__	$$a	ALEPH experiment: Candidate of Higgs boson production	✎ ✗
246_1	$$a	Expérience ALEPH: Candidat de la production d'un boson Higgs	✎ ✗
260__	$$c	14 05 2000	✎ ✗
340	$$a	FILM	✎ ✗
520__	$$a	Candidate for the associated production of the Higgs boson and Z boson. Both, the Higgs and Z boson decay into 2 jets each. The green and the yellow jets belong to the Higgs boson. They represent the fragmentation of a bottom andanti-bottom quark. The red and the blue jets stem from the decay of the Z boson into a quark anti-quark pair. Left: View of the event along the beam axis. Bottom right: Zoom around the interaction point at the centre showing detailsof the fragmentation of the bottom and anti-bottom quarks. As expected for b quarks, in each jet the decay of a long-lived B meson is visible. Top right: "World map" showing the spatial distribution of the jets in the event.	✎ ✗
595	$$a	Press	✎ ✗
65017	$$2 $$a	SzGeCERN Experiments and Tracks	✎ ✗
6531_	$$a	LEP	✎ ✗
8560_	$$f	neil.calder@cern.ch	✎ ✗



Record Editor: Record #6 (view history)

□ 001		6
□ 037__	□ $$a	CERN-DI-6206002
□ 041	□ $$a	eng
□ 245__	□ $$a	At CERN in 1962
	□ $$s	eight Nobel prizewinners
□ 260__	□ $$c	1962
☑ 520	☑ $$a	In 1962, CERN hosted the 11th International Conference on High Energy Physics. Among the distinguished visitors were eight Nobel prizewinners.Left to right: Cecil F. Powell, Isidor I. Rabi, Werner Heisenberg, Edwin M. McMillan, Emile Segre, Tsung Dao Lee, Chen Ning Yang and Robert Hofstadter.
□ 590__	□ $$a	En 1962, le CERN est l'hote de la onzieme Conference Internationale de Physique des Hautes Energies. Parmi les visiteurs eminents se trouvaient huit laureats du prix Nobel.De gauche a droite: Cecil F. Powell, Isidor I. Rabi, Werner Heisenberg, Edwin M. McMillan, Emile Segre, Tsung Dao Lee, Chen Ning Yang et Robert Hofstadter
□ 595__	□ $$a	Press
□ 65017	□ $$2	SzGeCERN
	□ $$a	Personalities and History of CERN
□ 6531_	□ $$a	Nobel laureate
□ 8564_	□ $$u	http://pcsimkoa9.cern.ch/record/6/files/6206002.jpg
□ 8564_	□ $$q	http://pcsimkoa9.cern.ch/record/6/files/icon-6206002.gif
	□ $$x	icon
□ 909C0	□ $$o	0000736PHOPHO
□ 909C0	□ $$y	1962
□ 909C0	□ $$b	81

- BibCheck module: allows to run quality checking on the metadata (command line)
- BibKnowledge module: allows to maintain Authority files and apply them to set of records
- BibExport module: allow to run batch export of records in many formats, for manual corrections and re-upload
- Holding pen: this is the place where the pending treatments to perform to records are listed
- BibMerge module: allows to merge identical records, preserving the information of the two records



Record Merger

Record1: 2

Record2: 148

Method: Revisions

Field	Subfield	Record 1	Record 2
001		2	148
037__	$a	CERN-EX-0104007	CERN-EX-0104007
100	$a		Patrice Loiez
245__	$a	The first CERN-built module of the barrel section of ATLAS's electromagnetic calorimeter	The first CERN-built module of the barrel section of ATLAS's electromagnetic calorimeter
246_1	$a	Premier module du tonneau du calorimètre électromagnétique d'ATLAS	
260__	$c	10 Apr 2001	12 Apr 2001

Conclusion

- In Invenio, a record is described in XMLMARC, and it can be associated with multiple files
- A record moves in the system through multiple modules, where it can be enriched with more information
- Submission of documents by (or in the name of) authors can be configured in multiple ways
- Records can also be imported from other systems in many different ways, including fulltexts
- Librarians have powerful tools to edit and improve existing records