

Ingestion via BatchUploader

Invenio User Group Workshop
7-9 May 2012 CERN

Javier Martin Montull <javier.martin.montull@cern.ch>
INSPIRE developer

Contents

- Feature overview
- Using the web interface
 - ▶ Uploading metadata
 - ▶ Uploading documents
- Daemon mode
- Robot upload

Feature overview

- Simplifies admin tools (bibupload, bibdocfile) for librarian usage
- Possibility to upload MarcXML and TextMARC
- Select upload mode (e.g. insert, replace) and schedule it for any given time
- Upload multiple files from a folder and perform the matching with existing records
- View personal upload history
- 'Daemon mode' that looks for new files periodically
- 'RobotUpload' facility to interface with systems through command line

Using the web interface

Metadata batch upload

Menu

0. Metadata batch upload 1. [Document batch upload](#) 2. [Upload history](#) 3. [Daemon monitor](#)

* Select file to upload: No file chosen

* File type:

* Upload mode: ⓘ

Upload priority:

Strong tags:

Upload later? then select: Date: Time: Example: 2009-12-20 19:22:18

*All fields with * are mandatory*

<https://inspireheptest.cern.ch/batchuploader>

Demo (metadata)

Select a file to upload and its type

Metadata batch upload

Menu

0. Metadata batch upload 1. [Document batch upload](#) 2. [Upload history](#) 3. [Daemon monitor](#)

* Select file to upload: correct_marxml.xml
* File type: MarcXML
 TextMARC
* Upload mode:

Supported types are:

MarcXML

TextMARC

```
▼<record>  
  <controlfield tag="001">1113526</controlfield>  
  <controlfield tag="005">20120505222029.0</controlfield>  
  ▼<datafield tag="970" ind1=" " ind2=" ">  
    <subfield code="a">SPIRES-9582649</subfield>  
  </datafield>
```

```
001113526 0247_ $$2DOI$$a10.1088/1748-0221/6/01/C01053  
001113526 037_ $$aSLAC-REPRINT-2012-012  
001113526 035_ $$9SPIRESTeX$$zGraf:2010zz  
001113526 100_ $$aGraf, N.A.$$uSLAC
```

<http://www.loc.gov/marc/bibliographic/ecbdhome.html>

Demo (II)

Metadata batch upload

Menu

0. Metadata batch upload 1. [Document batch upload](#) 2. [Upload history](#) 3. [Daemon monitor](#)

* Select file to upload: correct_marcxml.xml

* File type:

* Upload mode:

Upload priority:

Strong tags:

Upload later? then select

*All fields with * are mandatory*

Select an upload mode:

<https://inspireheptest.cern.ch/help/admin/bibupload-admin-guide#3.3>

Demo (III)

Metadata batch upload

Menu

0. Metadata batch upload 1. [Document batch upload](#) 2. [Upload history](#) 3. [Daemon monitor](#)

* Select file to upload: correct_marxml.xml

* File type:

* Upload mode: 

Upload priority: normal
 high

Strong tags:

The bibupload task that will go into BibSched can have priority:

normal - priority 1

high - priority 5

Demo (IV)

Metadata batch upload

Menu

0. Metadata batch upload 1. [Document batch upload](#) 2. [Upload history](#) 3. [Daemon monitor](#)

* Select file to upload: correct_marcxml.xml

* File type:

* Upload mode: ⓘ

Upload priority:

Strong tags:

Upload later? then select

Invenio configuration allows to set some tags as non replaceable. E.g. used in workflows where some content comes from a different system. Avoids replacing the references on INSPIRE when content comes from old Spire

Demo (V)

Metadata batch upload

Menu

0. Metadata batch upload 1. [Document batch upload](#) 2. [Upload history](#) 3. [Daemon monitor](#)

* Select file to upload: correct_marxml.xml

* File type:

* Upload mode: 

Upload priority:

Strong tags:

Upload later? then select: Example: 2009-12-20 19:22:18

All fields with * are mandatory

May 2012						
Su	Mo	Tu	We	Th	Fr	Sa
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

Program the task for a given time

An upload simulation is performed, and a summary of the actions to be performed displayed

Confirm your actions

No errors were found during the upload simulation

You are about to submit a **marcxml** file with name **C:\fakepath\correct_marcxml.xml** and content:

```
<?xml version="1.0" encoding="UTF-8"?>
<collection xmlns="http://www.loc.gov/MARC21/slim">
  <record>
    <controlfield tag="001">1111731</controlfield>
    <controlfield tag="005">20120421123229.0</controlfield>
    <datafield tag="024" ind1="7" ind2=" ">
      <subfield code="2">DOI</subfield>
      <subfield code="a">10.1103/PhysRevE.79.060105</subfield>
    </datafield>
    <datafield tag="035" ind1=" " ind2=" ">
      <subfield code="9">SPIRETeX</subfield>
      <subfield code="z">Enciso:2009zz</subfield>
    </datafield>
    <datafield tag="100" ind1=" " ind2=" ">
      <subfield code="a">Enciso, Alberto</subfield>
      <subfield code="u">Zurich, ETH D-Math</subfield>
    </datafield>
    <datafield tag="245" ind1=" " ind2=" ">
      <subfield code="a">Spin chains of Haldane-Shastry type and a generalized central limit
theorem</subfield>

```

This file will be uploaded with priority **Normal** and in mode **--correct**.

Do you want to submit the changes? (1 record(s) will be affected)

Confirm

Cancel

If there is a problem with the file to be uploaded, it will be displayed on the confirmation step

Confirm your actions

Some errors have been found during the upload simulation

Error: MARCXML file has wrong format: [(None, 0, 'Error: Mismatched end tag: expected , got \n in unnamed entity at line 51 char 13 of string input\nMismatched end tag: expected , got \nParse Failed!\n')]

You are about to submit a **marcxml** file with name **C:\fakepath\incorrect_closetag.xml** and content:

```
<?xml version="1.0" encoding="UTF-8"?>
<collection xmlns="http://www.loc.gov/MARC21/slim">
  <record>
    <controlfield tag="001">1111731</controlfield>
    <controlfield tag="005">20120421123229.0</controlfield>
    <datafield tag="024" ind1="7" ind2=" ">
      <subfield code="2">DOI</subfield>
      <subfield code="a">10.1103/PhysRevE.79.060105</subfield>
    </datafield>
    <datafield tag="035" ind1=" " ind2=" ">
      <subfield code="9">SPIRETeX</subfield>
      <subfield code="z">Enciso:2009zz</subfield>
    </datafield>
    <datafield tag="100" ind1=" " ind2=" ">
      <subfield code="a">Enciso, Alberto</subfield>
      <subfield code="u">Zurich, ETH D-Math</subfield>
    </datafield>
    <datafield tag="245" ind1=" " ind2=" ">
      <subfield code="a">Spin chains of Haldane-Shastry type and a generalized central limit
theorem</subfield>
```

This file will be uploaded with priority **Normal** and in mode **--insert**.

Do you want to submit the changes? (1 record(s) will be affected)

Confirm

Cancel

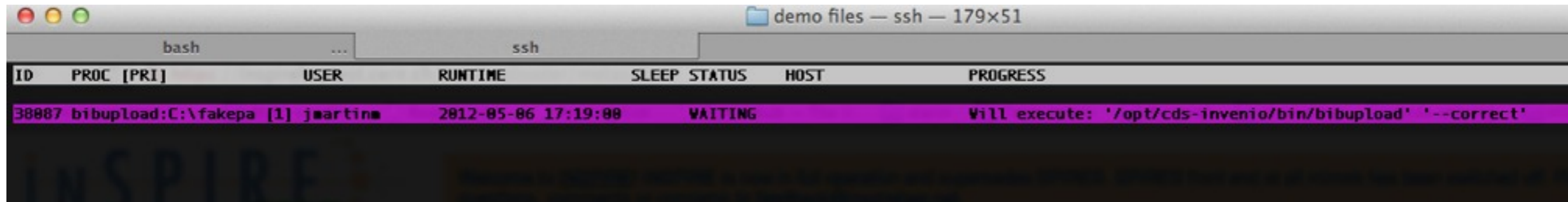
finally the upload task is added to BibSched

Upload successful

Menu

0. [Metadata batch upload](#) 1. [Document batch upload](#) 2. [Upload history](#) 3. [Daemon monitor](#)

Your file has been successfully queued. You can check your [upload history](#) or [submit another file](#)



The image shows a terminal window titled "demo files — ssh — 179x51". The terminal displays a table of tasks in a queue. The table has columns for ID, PROC [PRI], USER, RUNTIME, SLEEP, STATUS, HOST, and PROGRESS. A single task is listed with ID 38887, PROC bibupload, USER jmartin, and STATUS WAITING. The PROGRESS column indicates the command to be executed: "/opt/cds-invenio/bin/bibupload' '--correct'".

ID	PROC [PRI]	USER	RUNTIME	SLEEP	STATUS	HOST	PROGRESS
38887	bibupload	jmartin	2012-05-06 17:19:00		WAITING		Will execute: '/opt/cds-invenio/bin/bibupload' '--correct'

the actions performed are visible in the history tab

Upload history

Menu

- 0. [Metadata batch upload](#)
- 1. [Document batch upload](#)
- 2. Upload history
- 3. [Daemon monitor](#)

METADATA UPLOADS

Submit time	File name	Execution time	Status
2012-05-06 17:19:00	C:\fakepath\correct_marcxml.xml	2012-05-06 17:19:00	WAITING
2012-05-06 17:09:51	C:\fakepath\correct_marcxml.xml	2012-05-06 17:09:51	DONE

DOCUMENT UPLOADS

No document files have been uploaded yet.

Uploading documents

Document batch upload

Menu

0. [Metadata batch upload](#) 1. Document batch upload 2. [Upload history](#) 3. [Daemon monitor](#)

* Input directory: *Example: /afs/cern.ch/user/j/user/public/foo/*

* Filename matching:

* Upload mode: append revise

Upload priority:

Upload later? then select: *Date:* *Time:* *Example: 2009-12-20 19:22:18*

*All fields with * are mandatory*

Upload

Uploading documents

Document batch upload

Menu

0. [Metadata batch upload](#) 1. Document batch upload 2. [Upload history](#) 3. [Daemon monitor](#)

* Input directory: Example: [/afs/cern.ch/user/j/user/public/fool](#)

The directory has to be accessible from the server running Invenio



Uploading documents

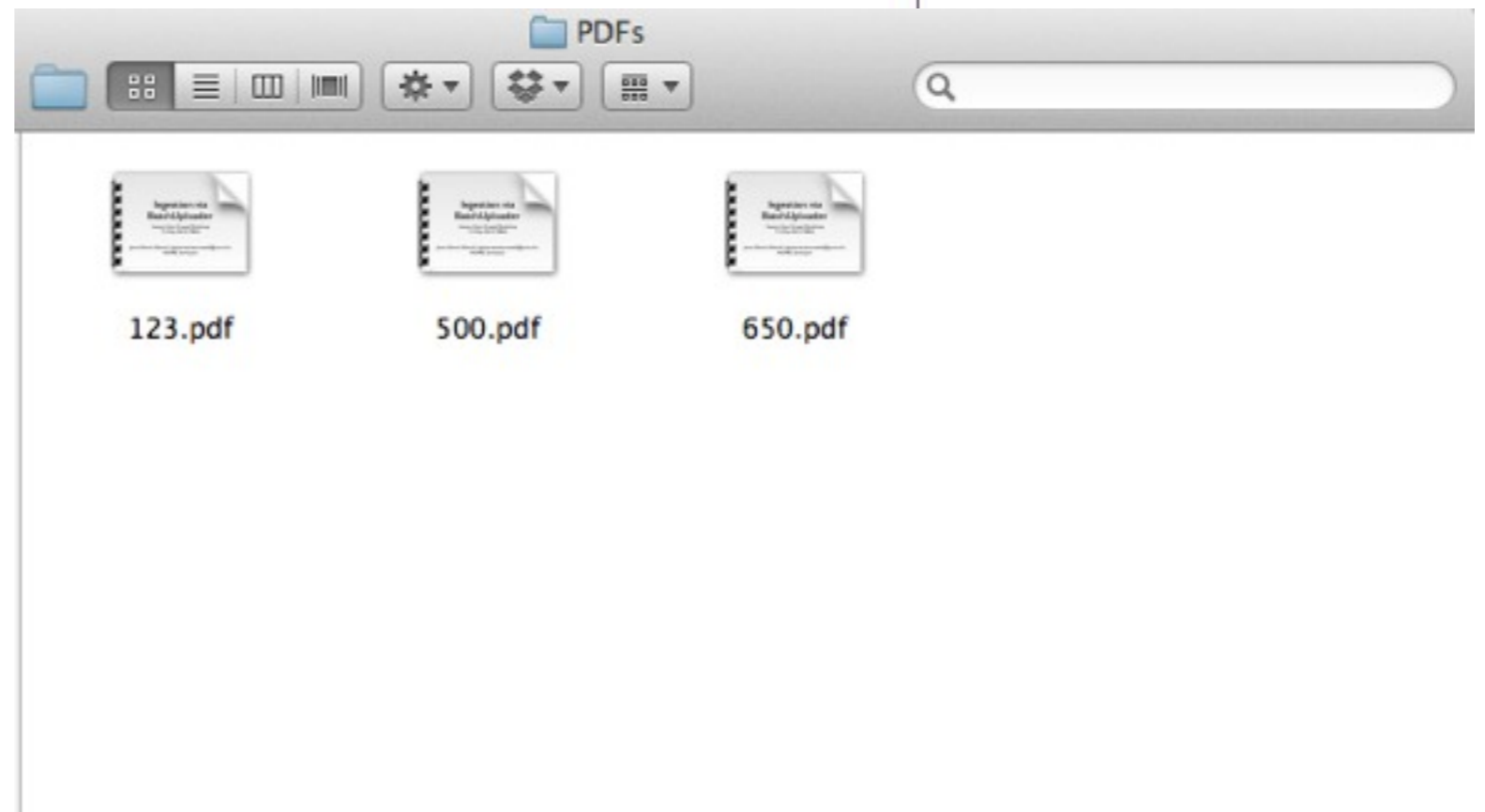
Document batch upload

Menu

0. [Metadata batch upload](#) 1. Document batch upload 2. [Upload history](#) 3. [Daemon monitor](#)

* Input directory: Example: */afs/cern.ch/user/j/user/public/foo/*

* Filename matching reportnumber
 recid



In order to determine the record the PDF belongs to, use either recid or reportnumber as the file name

Uploading documents

Document batch upload

Menu

0. [Metadata batch upload](#) 1. Document batch upload 2. [Upload history](#) 3. [Daemon monitor](#)

* Input directory: *Example: /afs/cern.ch/user/j/user/public/fool*
* Filename matching:
* Upload mode: append revise
Upload priority:

<https://inspireheptest.cern.ch/help/admin/bibupload-admin-guide#3.6>

Priority inside of BibSched queue

Uploading documents

Get feedback on actions performed

Document batch upload result

Menu

0. [Metadata batch upload](#) 1. [Document batch upload](#) 2. [Upload history](#) 3. [Daemon monitor](#)

3 documents have been found.

The following files have been successfully queued:

- **123.pdf**
- **500.pdf**
- **650.pdf**

The following errors have occurred:

Some files could not be moved to DONE folder. Please remove them manually.

[Return to upload form](#)

Daemon mode

- Intended to be a BibSched task for periodical document or metadata upload
- The parent directory where the daemon will look for folders *metadata* and *documents* must be specified in the invenio configuration file.
- When running the batchuploader daemon there are two possible execution modes:
 - ▶ `-m, --metadata` Look for metadata files in folders insert, append, correct and replace.
 - ▶ `-d, --documents` Look for documents in folders append and revise.
- Example:
 - ▶ `$ batchuploader --documents`

Batch Uploader: Daemon monitor

Menu

0. [Metadata batch upload](#) 1. [Document batch upload](#) 2. [Upload history](#) 3. Daemon monitor

- Using **web interface upload**, actions are executed a single time.
- Check the [Batch Uploader daemon help page](#) for executing these actions periodically.

METADATA FOLDERS

- **/opt/cds-invenio/var/batchupload/metadata/replace**
- **/opt/cds-invenio/var/batchupload/metadata/insert**
www-data www-data 9 2010-06-21 08:33 new_insertions.xml
- **/opt/cds-invenio/var/batchupload/metadata/correct**
- **/opt/cds-invenio/var/batchupload/metadata/append**
www-data www-data 13 2010-06-21 08:33 test_marc.xml
www-data www-data 10 2010-06-21 08:32 marc_corrections.xml

DOCUMENT FOLDERS

- **/opt/cds-invenio/var/batchupload/documents/append**
www-data www-data 11 2010-06-21 08:34 CERN-PRE-82-006.pdf
- **/opt/cds-invenio/var/batchupload/documents/revise**

Last BibSched tasks:

ID	Name	Time	Status	Progress
	batchuploader			Not executed yet

Next scheduled BibSched run:

ID	Name	Time	Status	Progress
	batchuploader			Not scheduled

Contents

- Feature overview
- Using the web interface
 - ▶ Uploading metadata
 - ▶ Uploading documents
- Daemon mode
- Robot upload

Questions?

javier.martin.montull@cern.ch