Invenio v1.2

Tibor Šimko

CERN

Invenio User Group Workshop 2013
Jülich, Germany

18–20 November 2013

» Outline

1

Overview

» Invenio v1.2 (WIP)

- history:
  - work-in-progress
  - to be released in 2013
- status:
  - better restricted material search
  - better circulation module
  - new Solr/Xapian ranking bridge
  - new task schedule monitor
  - new record upload differ
  - new indexer taking advantage of record revisions
  - new virtual global index
  - new non-MARC master format support
  - new JSON record structure
  - new search API
  - new DOI minter (via DataCite)

2

Optimising Search Experience

INVENIO › Announcing News

INVENIO » Add-To-Search UI

CERN Accelerating science

CERN Document Server
Access articles, reports and multimedia content in HEP

Search | Submit | Help | Personalize

Search 1,315,834 records for:

[   ] Search   Search Tips
                                              Advanced Search

Add to Search −

[AND ▼] [All of the words: ▼] [   ] [any field ▼] [Add to Search]

☐ Search also in the full-text of all documents

☑ Articles & Preprints (1,116,490)

Published Articles (342,699) Preprints (689,341)
Theses (19,300) Reports (5,809)

CERN Articles & Preprints (107,222)

CERN Published Articles (58,460) CERN Preprints (17,898) CERN Theses (5,083)
CERN Reports (1,207) Committee Documents (24,692)

INVENIO » Search Services

- teasers to display things, e.g. author profiles, e.g. help messages

» Searching for Restricted

- no need to pre-selected restricted collections prior to search

admin   guest

3

Optimising Ingestion

» Problematic Uploader

- replace mode replaces everything → too many SQL queries
- example: input file 303 records, changes in 980 field only

```
 ncalls   tottime  percall  cumtime   percall  filename:lineno(function)
      1    0.002    0.002   385.160   385.160  bibtask.py:709(_task_run)
      1    0.038    0.038   385.127   385.127  bibupload.py:2037(task_run_core)
    303    0.065    0.000   384.588     1.269  bibupload.py:125(bibupload)
 153403    1.997    0.000   337.916     0.002  dbquery.py:151(run_sql)
 153403    2.956    0.000   331.396     0.002  cursors.py:129(execute)
 153403    0.543    0.000   325.573     0.002  cursors.py:311(_query)
 153403  319.412    0.002   323.033     0.002  cursors.py:273(_do_query)
    303    0.557    0.002   187.090     0.617  bibupload.py:1617(update_databas
  43601    0.350    0.000   117.853     0.003  bibupload.py:890(insert_record_b
    303    0.057    0.000   103.269     0.341  bibupload.py:1847(delete_bibrec_
    239    0.296    0.001    72.473     0.303  bibupload.py:910(synchronize_856
  43601    1.406    0.000    67.998     0.002  bibupload.py:845(insert_record_b
```

- executed 150k SQL queries (!) to update all bibxxx MARC tables
- what if backend behaved more like "git diff" and "git patch"?

INVENIO » Smarter Uploader

- smarter uploader:
  1. detect changed fields by comparing revision numbers (tag 005)
  2. update only necessary bibxxx MARC tables → **few SQL queries**
  3. log which fields are affected by given change:

recID	revision	affected_fields
123	2013-12-01 11:00:23	100,700
123	2013-12-02 17:15:44	035,245,999
123	2013-12-08 03:01:18	100
123	2013-12-08 03:01:18	firstauthor

- record differ/patcher lessens need for record locking
  – John and Jane can edit the same record at the same time
  – upload backend figures out wanted changes from each cataloguer
  – auto-merge wanted changes if possible
  – alert human curator if not possible

INVENIO » Smarter Indexer

- single last_updated date per record → reindex all indexes if one field changes
  - safety net, but time consuming
- take advantage of affected_fields of every pending record
  - if title changes, author etc indexes are not updated
  - one **record +**
    one single **last_modified** date **+**
    logging of **affected_fields** per revision upload **=**
    multiple last_modified dates per field
- technique to be spread to BibRank, BibSort, BibAuthorID etc daemons

» Virtual Global Index

- **global index: re-tokenises everything**
  - time-consuming in case of paper claiming and author IDs
- **what if indexer amended global fields while indexing given index?**
  - notion of "virtual" vs "normal" index
  - virtual updated only via normal indexes

ID	Virtual index	Dependent indexes
1	global	collection, abstract, keyword, reportnumber, title, year, journal, collaboration, affiliation, exacttitle, miscellaneous

  - need of a "miscellaneous" index

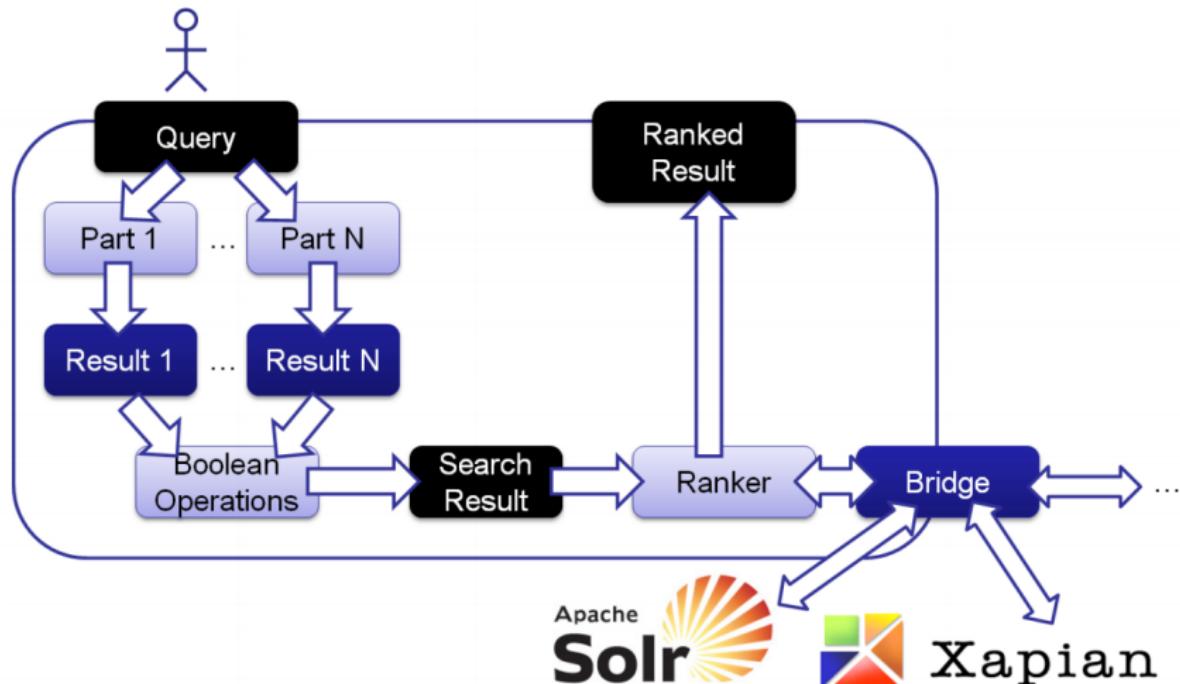|	
miscellaneous	01%, 02%, 031%, 032%, 033%, 034%, 035%, 036%, 037%, 038%, 04%, 05%, 06%, 07%, 080%, 082%, 083%, 084%, 085%, 086%, 09%, 100__0, 100__a, 11%, 110__0, 12%, 13%, 130__0, 14%, 15%, 150__0, 16%, 17%, 18%, 19%, 20%, 21%, 22%, 23%, 240%, 242%, 243%, 244%, 247%, 25%, 26%, 260__0, 27%, 28%, 29%, 30%, 31%, 32%, 33%, 34%, 35%, 36%, 37%, 38%, 39%, 40%, 41%, 42%, 43%, 44%, 45%, 46%, 47%, 48%, 49%, 50%, 51%, 521%, 522%, 524%, 525%, 526%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 650%, 651%, 6531_9, 6531_v, 6531_y, 654%, 655%, 656%, 657%, 658%, 66%, 67%, 68%, 69%, 700__0, 700__a, 711%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 850%, 851%, 852%, 853%, 854%, 855%, 857%, 858%, 859%, 86%, 87%, 88%, 89%, 900%, 901%, 902%, 903%, 904%, 905%, 906%, 907%, 908%, 909C1%, 909C5%, 909CK%, 909CO%, 909CP%, 909CS%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 981%, 982%, 983%, 984%, 985%, 986%, 987%, 988%, 989%

INVENIO » Configurability of Indexes

■ centralisation of index configurations: stemming, synonyms, stopwords, HTML/LaTeX markup, tokeniser, type

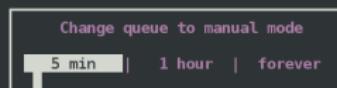| ID	Name	Fwd.Idx Size	Rev.Idx Size	Fwd.Idx Words	Rev.Idx Records	Last updated	Fields	Translations	Stemming Language	Synonym knowledge base	Remove stopwords	Remove HTML markup	Remove Latex markup	Tokenizer	Indexer type
1	global	0.238 MB	0.171 MB	6,632	141	Not updated	any field	None	english	INDEX-SYNONYM-TITLE.exact	No	No	No	BibIndexDefaultTokenizer	virtual
2	collection	0.001 MB	0.009 MB	36	141	2013-11-17 05:34:56	collection	None	None	None	No	No	No	BibIndexDefaultTokenizer	native
3	abstract	0.123 MB	0.043 MB	3,516	80	2013-11-17 05:34:56	abstract	None	english	None	No	No	No	BibIndexDefaultTokenizer	native
4	author	0.024 MB	0.013 MB	722	141	2013-11-17 05:34:56	author	None	None	None	No	No	No	BibIndexAuthorTokenizer	native
5	keyword	0.001 MB	0.001 MB	27	9	2013-11-17 05:34:56	keyword	None	english	None	No	No	No	BibIndexDefaultTokenizer	native
6	reference	0.253 MB	0.075 MB	7,262	48	2013-11-17 05:34:56	reference	None	None	None	No	No	No	BibIndexDefaultTokenizer	native
7	reportnumber	0.013 MB	0.008 MB	374	89	2013-11-17 05:34:56	report number	None	None	None	No	No	No	BibIndexDefaultTokenizer	native
8	title	0.022 MB	0.013 MB	665	115	2013-11-17 05:34:56	title	None	english	INDEX-SYNONYM-TITLE.exact	No	No	No	BibIndexDefaultTokenizer	native
9	fulltext	1.428 MB	0.42 MB	39,366	73	2013-11-17 05:34:56	fulltext	None	english	None	No	No	No	BibIndexFulltextTokenizer	native
10	year	0.001 MB	0.004 MB	30	78	2013-11-17 05:34:56	year	None	None	None	No	No	No	BibIndexYearTokenizer	native
11	journal	0.006 MB	0.009 MB	134	141	2013-11-17 05:34:56	journal	None	None	None	No	No	No	BibIndexJournalTokenizer	native
12	collaboration	0.001 MB	0.001 MB	32	5	2013-11-17 05:34:56	collaboration	None	None	None	No	No	No	BibIndexDefaultTokenizer	native
13	affiliation	0.004 MB	0.007 MB	126	79	2013-11-17 05:34:56	affiliation	None	None	None	No	No	No	BibIndexDefaultTokenizer	native
14	exactauthor	0.024 MB	0.013 MB	722	141	2013-11-17 05:34:56	exact author	None	None	None	No	No	No	BibIndexExactAuthorTokenizer	native

INVENIO » Using external rankers

- taking advantage of external ranking systems
- Solr, Xapian supported; ElasticSearch to come?

» Smarter Task Scheduler

- better task parallelism and auto execution
- auto-wakeup from manual mode

» Ingestion Health Monitor

```
Records pending per indexing method since ........................................ N/A
  - global ................................................................. 141
  - collection ............................................................... 4
  - abstract ................................................................. 4
  - author ................................................................... 4
  - keyword ........................................................... Up to date
  - reference ................................................................ 4
  - reportnumber ............................................................. 4
  - title ............................................................. Up to date
  - fulltext ................................................................. 4
  - year ..................................................................... 4
  - journal .................................................................. 4
  - collaboration ............................................................ 4
  - affiliation .............................................................. 4
  - exactauthor .............................................................. 4
  - caption .................................................................. 4
  - firstauthor .............................................................. 4
  - exactfirstauthor ......................................................... 4
  - authorcount .............................................................. 4

Records pending per ranking method since .......................... 17 Nov 2013, 05:40
  - wrd ...................................................................... 4
  - demo_jif ................................................................. 4
  - citation ................................................................. 4
  - citerank_citation_t ...................................................... 4
  - citerank_pagerank_c ...................................................... 4
  - citerank_pagerank_t ...................................................... 4

Records pending per sorting method since .......................... 17 Nov 2013, 05:40
  - latest first ............................................................. 4
  - title .................................................................... 4
  - author ................................................................... 4
```

Authorities

INVENIO) » Authorities

| Search | Submit | Personalize ▾ | Help | Administration ▾ |

Home > Authority Records

Authority Records

Search 24 records for:

[   ] [any field ▾] **Search** **Browse**

Search Tips :: Advanced Search

Narrow by collection:

- ☑ Authors (5)
- ☑ Institutions (15)
- ☑ Journals (0)
- ☑ Subjects (4)

INVENIO » Authority Record

ATLANTIS INSTITUTE OF FICTIVE SCIENCE

| Search | Submit | Personalize ▼ | Help | Administration ▼ |

Home > Authority Records > Authors > Record #118

| Information | References (0) | Citations (0) | Keywords | Discussion (0) | Usage statistics | Files | Plot |
| Linkbacks |

Authority Record

Control Number(s): AUTHOR|(DLC)n 80141717 (0 dependent records), AUTHOR|(SzGeCERN)aaa0005 (4 dependent records)

Main author name: Ellis, John, 1946-

Variant(s): Ellis, J., 1946-, Ellis, Jonathan Richard, 1946-

ATLANTIS INSTITUTE OF FICTIVE SCIENCE

Search | Submit | Personalize ▾ | Help | Administration ▾

Home > Authority Records > Record #118 > MARC

```
000000118 001__ 118
000000118 005__ 20131117053442.0
000000118 035__ $$aAUTHOR|(DLC)n 80141717
000000118 035__ $$aAUTHOR|(SzGeCERN)aaa0005
000000118 100__ $$aEllis, John$$d1946-
000000118 400__ $$aEllis, J.$$d1946-$$q(John),
000000118 400__ $$aEllis, Jonathan Richard$$d1946-
000000118 670__ $$aEurophysics Study Conference on Unification of the Funda
000000118 670__ $$aSupersymmetry and supergravity, c1986:$$bCIP t.p. (J. El
000000118 670__ $$aQuantum reflections, 2000:$$bCIP t.p. (John Ellis) data
000000118 980__ $$aAUTHORITY
000000118 980__ $$aAUTHOR
```

INVENIO » Bibliographic Record

```
000000018 001__ 18
000000018 005__ 20131117053300.0
000000018 041__ $$aeng
000000018 088__ $$aCERN-TH-4036
000000018 100__ $$0AUTHOR|(SzGeCERN)aaa0005$$0INSTITUTION|(SzGeCERN
000000018 245__ $$aNon-compact supergravity solves problems
000000018 260__ $$aGeneva$$bCERN$$cOct 1984
000000018 300__ $$a15 p
000000018 65017 $$2SzGeCERN$$aGeneral Theoretical Physics
000000018 6531_ $$aKahler
000000018 6531_ $$amanifolds
000000018 6531_ $$agravitinos
000000018 6531_ $$aaxions
```

INVENIO » Bibliographic Search

ATLANTIS INSTITUTE OF FICTIVE SCIENCE

Search | Submit | Personalize ▾ | Help | Administration ▾

Home > Search Results: author:"Ellis, Jonathan Richard"

Search:

author:"Ellis, Jonathan Richard"   any field ▾   Search   Browse

Search Tips :: Advanced Search

Search collections:

*** any public collection *** ▾

Sort by: **Display results:**   **Output format:**

latest first ▾ desc. ▾ - or rank by - ▾   10 results ▾ split by collection ▾   HTML brief ▾

Results overview: Found **4** records in 0.07 seconds.

Articles & Preprints, **4 records found**

Articles & Preprints   **4** records found

☐ 1. **Non-compact supergravity solves problems** / Ellis, J ; Enqvist, K ; Nanopoulos, D V [CERN-TH-4036]
   Published in **Phys. Lett., B :151 1985 357-362**
   Detailed record - Similar records - Cited by 1 record

☐ 2. **From the standard model to grand unification** / Ellis, J [CERN-PRE-82-006]
   Detailed record - Similar records

☐ 3. **Physics at the front-end of a neutrino factory : a quantitative appraisal** / Mangano, M L ; Alekhin, S I ; Anselmino
   [hep-ph/0105155]
   We present a quantitative appraisal of the physics potential for neutrino experiments at the front-end of a muon storage ring. [...]
   Fulltext: PDF PS.GZ;
   Detailed record - Similar records

☐ 4. **Grand unification with large supersymmetry breaking** / Ellis, J ; Ibanez, L E ; Ross, G G [PRE-25553] [RL-82-024]

5

JSON-ification

» Non-MARC Master Formats

- non-MARC master formats: UNIMARC, EAD
- standardising into JSON
- testing speed: MongoDB vs PostreSQL
- extending JSON store for non-record namespaces
- search API

» Record JSON Configuration

```
authors[0], creator:
    creator:
        @legacy((("100", "100__", "100__%"), ""),
                ("100__a", "first author name", "full_name"),
                ("100__e", "relator_name"),
                ("100__h", "CCID"),
                ("100__i", "INSPIRE_number"),
                ("100__u", "first author affiliation", "affiliation"))
        marc, "100__", { 'full_name':value['a'], 'first_name':util_split(value['a']
    checker:
        check_field_existence(0,1)
        check_field_type('str')
    documentation:
        "Main Author"
    producer:
        json_for_marc, {"100__a": "full_name", "100__e": "relator_name", "100__h":
        json_for_dc, {"dc:creator": "full_name"}

_number_of_comments:
    calculated:
        @parse_first(('recid', ))
        get_number_of_comments(self.get('recid'))
    documentation:
```

» Record in JSON

```
In [1]: from invenio.bibfield import get_record
In [2]: rec = get_record(81)

In [3]: rec['authors']
Out[3]:
[{'affiliation': 'Stanford University',
  'first_name': 'A',
  'full_name': 'Adams, A',
  'last_name': 'Adams'},
 {'first_name': 'J',
  'full_name': 'McGreevy, J',
  'last_name': 'McGreevy'},
 {'first_name': 'E',
  'full_name': 'Silverstein, E',
  'last_name': 'Silverstein'}]

In [4]: rec['authors.full_name']
Out[4]: ['Adams, A', 'McGreevy, J', 'Silverstein, E']
```

6

More!

INVENIO » More!

- more topics covered elsewhere throughout this workshop

- release date?
  - v1.2 vs v1.3 vs ... vs v2.0