

# BibAuthorID

a brief overview

Samuele Carli

9 maggio 2012

# Contents

- 1 The problem
- 2 Algorithmic approximation
- 3 Data flow
- 4 Usage
- 5 Paper claiming

# Author identities...

...are messed up! (Authors, curators, publishers, goblins...)

Mr. Ellis is known as:

- Ellis, John R.
- Ellis, John
- Ellis, J.
- Ellis, J.R.
- Ellis, John R., (ed.)
- Ellis, John R., (Ed.)
- Ellis, Jonathan R.
- Ellis, John, (ed.)
- Ellis, John.R.
- Ellis, J., (ed.)
- Ellis, Jonathan Richard

And who is this Ellis, J., by the way? John, Joseph or Jaqueline?

# BibAuthorID: one way to tackle the problem

What is needed:

- a. Unambiguous, stable references to signatures on papers
- b. Stable storage of clusters along with other generic cluster-related info
- c. Algorithmic way to cluster signatures together
- d. Authoritative and simple way to correct and hint the algorithms [Claiming interface]
- e. Algorithmic (fast!) way to keep storage up to date: new, updated, modified, removed records

## a.: stable references to signatures on papers

- Strongly coupled to underlying system
- in INVENIO: needs to go down to MARC tables

RecordID (Rec)  $\leftrightarrow$  BibXXX entry (Ref)

BibRefRec = 100:1,2  $\rightarrow$  id=1 in bib10x with RecID=2

BibRefRec = 700:25,89  $\rightarrow$  id=25 in bib70x with RecID=89

That's the only reference we can safely store for which is assured some stability in time (**almost!**).

## b.: storage of clusters and info

### aidPERSONIDPAPERS:

(**personid**, **bibref\_table**, **bibref\_value**, **bibrec**, name, flag, lcul, last\_updated)

(481884,700,35092,61707,'Shufeldt, W.',1,0,2012-03-29 23:36:54)

(481884,100,12687,65478,'Shufeldt, Walter',1,0,2012-03-29 23:36:55)

- personid is a unique identifier for a Person, which defines the signatures clusters

### aidPERSONIDDATA:

(**personid**, **tag**, **data**, opt1, opt2, opt3)

(1524, canonical\_name, M.Ihl.1, 0, 0, NULL)

(1524, uid, 1853, 0, 0, NULL)

# Author disambiguation: TORTOISE

- Compare all signatures...

- Strings
- Affiliations
- Coauthors
- Keywords
- Collaborations
- InspireIDs
- Human input (claims)
- ...

...to build a huge comparison Matrix.

- Clusterize the Matrix!

...of course, the whole thing is highly optimized, comparisons and computations are reduced to minimum with smart preclustering, caching, memoizations, ...

- INSPIRE: 6M signatures on 1M records

1 Week on a Xeon X5650 @ 2.67GHz, 6 cores+HyperThreading, 24Gb RAM

Main bottlenecks: Signatures comparison, data retrieval, swap space

# Author disambiguation: data maintenance

Problems so far:

- One week cycle not acceptable on production machines
- Data changes constantly: new records, modified records...  
→ leads to data inconsistencies (stale BibRefRecs) which **MUST** be corrected ASAP
- Users want new data as soon as it is produced



# Author disambiguation: RABBIT

- Fast approximated attribution of new papers
  - purely on names strings: cached in aidPERSONIDPAPERS are cheap and fast to use
  - Errors in attribution will get corrected by next run of TORTOISE
- Fast correction of stale BibRefRecs wherever possible

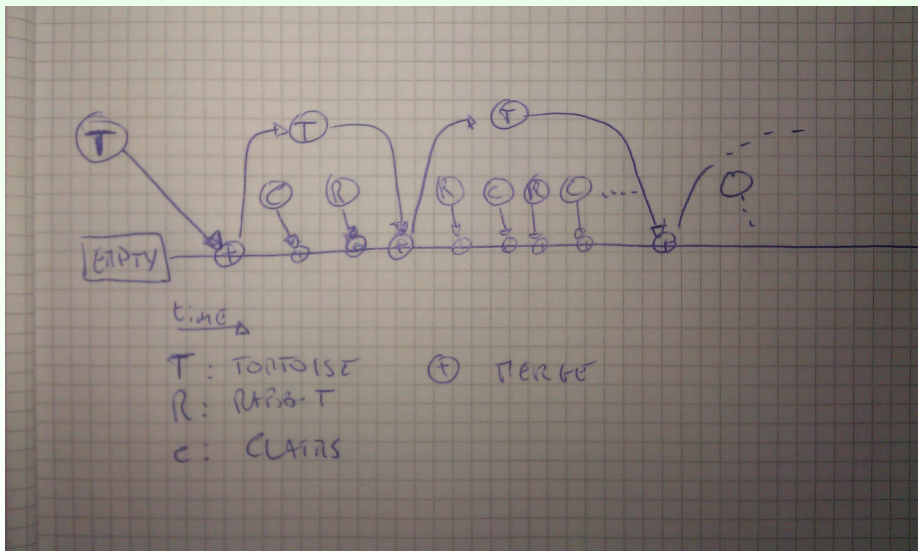
Stably in PRODUCTION on INSPIRE since March!

# Author disambiguation: MERGER

- Merges TORTOISE results with current aidPERSONIDPAPERS status
- Takes care of not touching authoritative entries (human input)
- Deals with one week old data from TORTOISE and bleeding edge new data on aidPERSONID: not trivial

Not in master yet, being tested!

# Data flow



At the moment the first filling can be done using RABBIT!

# Usage? really simple!

- install INVENIO
  - `$PREFIX/bin/bibauthorid --disambiguate`  
Be ready to wait...
  - `$PREFIX/bin/bibauthorid --merge`
  - `$PREFIX/bin/bibauthorid --update-personid`
    - scheduled regularly (on INSPIRE prod. runs every 12 hours)
    - runs only on modified or created records since last run
    - (INSPIRE plans to run it every hour)
- Ideally TORTOISE is run on a separate machine with lots of RAM and processors, then results are copied to a production machine and merged.
  
- Before MERGER gets in master, one can fully disambiguate starting with an empty `aidPERSONIDPAPERS` and:  
`$PREFIX/bin/bibauthorid --update-personid --all-records`

# Paper claiming

Allows:

- Manual correction of misclustered entries
- Manual suggestion of exceptions (marriage, name changes...)
- Confirmation of correct results: never get it wrong again!

# Authorpage: WebAuthorProfile (...soon on your screens!)

Available to users and curators, different permission levels (conflicts resolution)

## Albrow, M.G. (797 papers)

[This is me. Verify my publication list.](#)

### Name variants

Albrow, M.G. (537)  
 Albrow, Michael (109)  
 Albrow, Michael G. (105)  
 Albrow, M. (43)  
 Albrow, Michael G., (ed.) (2)  
 Albrow, M., (ed.) (1)

### Papers

	All papers	Single authored
<b>All papers</b>	<b>797</b>	<b>44</b>
Book	2	0
ConferencePaper	82	29
Introductory	4	3
Lectures	1	1
Preprint	0	0
Published	649	5
Review	21	10
Thesis	0	0

### Affiliations

Fermilab (557)  
 unknown affiliation (179)  
 Rutherford (33)  
 CERN (20)  
 Manchester U. (5)  
 Stockholm U. (3)  
 Argonne (1)  
 Pisa U. (1)  
 Hamburg U. (1)  
 Brookhaven (1)

### Collaborations

CDF Collaboration (408)  
 CMS Collaboration (128)  
 Axial Field Spectrometer Collaboration (47)  
 UA1 Collaboration (38)  
 CHLM Collaboration (9)  
 British-French-Scandinavian Collaboration (6)  
 CDF II Collaboration (6)  
 CDF - Run II Collaboration (4)  
 CDF - Run II (3)  
 The CDF Collaboration (3)  
[more](#)

### Frequent co-authors

[Rodrigo, Teresa](#) (573)  
[Barnes, Virgil E.](#) (556)  
[Laasanen, Alvin Toivo](#) (554)  
[Goulianos, Konstantin A.](#) (553)  
[Barnett, Bruce Arnold](#) (549)

### Frequent keywords

experimental results (532)  
 anti-p p: colliding beams (447)  
 CDF (434)  
 Batavia TEVATRON Coll (425)  
 anti-p p: inclusive reaction (359)  
 1800 GeV-cms (208)  
 1960 GeV-cms (168)  
 colliding beams: anti-p p (129)

### Citations (from papers in INSPIRE):

#### Citation summary results

	All papers	Published only
<b>Total number of citable papers analyzed:</b>	<b>698</b>	<b>649</b>
<b>Total number of citations:</b>	0	0
<b>Average citations per paper:</b>	0.0	0.0
<b>Total number of citations excluding self-citations [2]</b>	0	0
<b>Average citations per paper excluding self-citations [2]</b>	0.0	0.0
<b>Breakdown of papers by citations:</b>		
Renowned papers (500+)	0	0
Famous papers (250-499)	0	0
Very well-known papers (100-249)	0	0
Well-known papers (50-99)	0	0
Known papers (10-49)	0	0
Less known papers (1-9)	0	0
Unknown papers (0)	698	649
<b>Additional Citation Metrics [2]</b>		
h-index [2]	0	0

### HepNames data

**Michael G. Albrow** ([Fermilab](#))  
[\[Publication list\]](#) [\[Google\]](#) [\[Students\]](#) [\[arXiv\]](#) [\[ADS\]](#)

PhD Advisor: [Murphy, Paul G.](#)  
 PhD Institution: [Manchester U.](#)  
 Email: [albrow@fnal.gov](mailto:albrow@fnal.gov)  
[michael.albrow@cern.ch](mailto:michael.albrow@cern.ch)  
 Field: HEP-EX  
 Experiment: [CERN-LHC-CMS](#), [FNAL-E-0830](#)  
 Author ID: M.G.Albrow.1

# Papers claiming interface

## Attribute papers for: M.G.Albrow.1

Navigation: Run paper attribution for another author

### Names variants:

Albrow, M. (44); Albrow, M. G. (537); Albrow, Michael (109); Albrow, Michael G. (107);

Papers (797)

Papers removed from this profile (0)

Select All | Select None | Invert Selection | Hide successful claims

Yes, those papers are by this person.
  No, those papers are not by this person
  Assign to other person

Search:

	Paper Short Info	Author Name	Affiliation	Date	Experiment	Actions
<input type="checkbox"/>	1. <a href="#">Search for radiative decays of neutralinos in <math>\sqrt{s} = 1.8</math> TeV CDF Collaboration (F. Abe (KEK, Tsukuba) et al.)</a>	Albrow, M.	Fermilab	1994-07	FNAL-E-0741	<input checked="" type="checkbox"/> Yes, this paper is by this person. <input checked="" type="checkbox"/> No, this paper is <i>not</i> by this person <input type="checkbox"/> Assign to another person
<input type="checkbox"/>	2. <a href="#">Hadronic shower simulation. Proceedings, Workshop, Batavia, USA, September 6-8, 2006</a> M. Albrow, (ed.), R. Raja, (ed.) (Fermilab).	Albrow, M.	Fermilab	2007	N.A.	<input checked="" type="checkbox"/> Marked as this person's paper <input checked="" type="checkbox"/> But it's <i>not</i> this person's paper. <input type="checkbox"/> Assign to another person
<input type="checkbox"/>	3. <a href="#">Development of picoseconds Time of Flight systems in Meson Test Beam Facility at Fermilab</a> A. Ronzhin, M. Albrow, M. Demarteau, S. Los (Fermilab), S. Malik (Rockefeller U.), S. Pronko, E. Ramberg (Fermilab), A. Zatserklyany (Puerto Rico U., Mayaguez).	Albrow, M.	Fermilab	2010-11	N.A.	<input checked="" type="checkbox"/> Marked as this person's paper <input checked="" type="checkbox"/> But it's <i>not</i> this person's paper. <input type="checkbox"/> Assign to another person
<input type="checkbox"/>	4. <a href="#">Test of timing properties of the Photek 240 PMT</a> A. Ronzhin, M. Albrow, M. Demarteau, S. Pronko, E. Ramberg (Fermilab), A. Zatserklyany (Puerto Rico U., Mayaguez).	Albrow, M.	Fermilab	2010-01	N.A.	<input checked="" type="checkbox"/> Marked as this person's paper <input checked="" type="checkbox"/> But it's <i>not</i> this person's paper. <input type="checkbox"/> Assign to another person
<input type="checkbox"/>	5. <a href="#">Tests of timing properties of silicon photomultipliers</a> A. Ronzhin, M. Albrow (Fermilab), K. Byrum (Argonne), M. Demarteau, S. Los (Fermilab), E. May (Argonne), A. Ramberg (Fermilab), J. Va'vra (SLAC), A. Zatserklyany (Puerto Rico U., Mayaguez).	Albrow, M.	Fermilab	2010-03	N.A.	<input checked="" type="checkbox"/> Marked as this person's paper <input checked="" type="checkbox"/> But it's <i>not</i> this person's paper. <input type="checkbox"/> Assign to another person
<input type="checkbox"/>	6. <a href="#">Beam Test of a Time-of-Flight Detector Prototype</a> J. Va'vra, D. W. G. S. Leith, B. Ratcliff (SLAC), E. Ramberg, M. Albrow, A.	Albrow, M.	Fermilab	2009	N.A.	<input checked="" type="checkbox"/> Marked as this person's paper

# Ticket review before commit

## Please review your actions

*Please provide your information*

Your first name:

Your last name:

Your eMail:

You may leave a comment (optional):

Continue claiming\*

Confirm these changes\*\*

Delete the entire request!

*Mark as their documents*

Search for radiative decays of neutralinos in  $p\bar{p}$  collisions at  $\sqrt{s} = 1.8$  TeV

Albrow, M. G. (Selected name on paper: Albrow, M.)



Cancel



*Mark as not their documents*

Nothing staged in this category

\* You can come back to this page later. Nothing will be lost.

\*\* Performs all requested changes. Changes subject to permission restrictions will be submitted to an operator for manual review.

Symbols legend:

-  The result of this request will be visible immediately
-  The result of this request will be visible immediately but we need your confirmation to do so for this paper has been manually claimed before



Want to know more?

## Q&A session!

For further information call 1-800-bibauthorid-help or:

**Samuele Carli**

E-mail: [scarli@cern.ch](mailto:scarli@cern.ch)

Web: [www.csspace.net](http://www.csspace.net)