ULB

Measuring ULB scholars output visibility: A quantitative assessment of Scopus metadata quality using Google Refine

Sébastien Droesbeke, Seth van Hooland, Max De Wilde and Isabelle Boydens
Université libre de Bruxelles, Belgium
{sdroesbe, svhoolan, madewild, iboydens}@ulb.ac.be

LIBER 40th Annual Conference – Barcelona
30 June 2011

Context

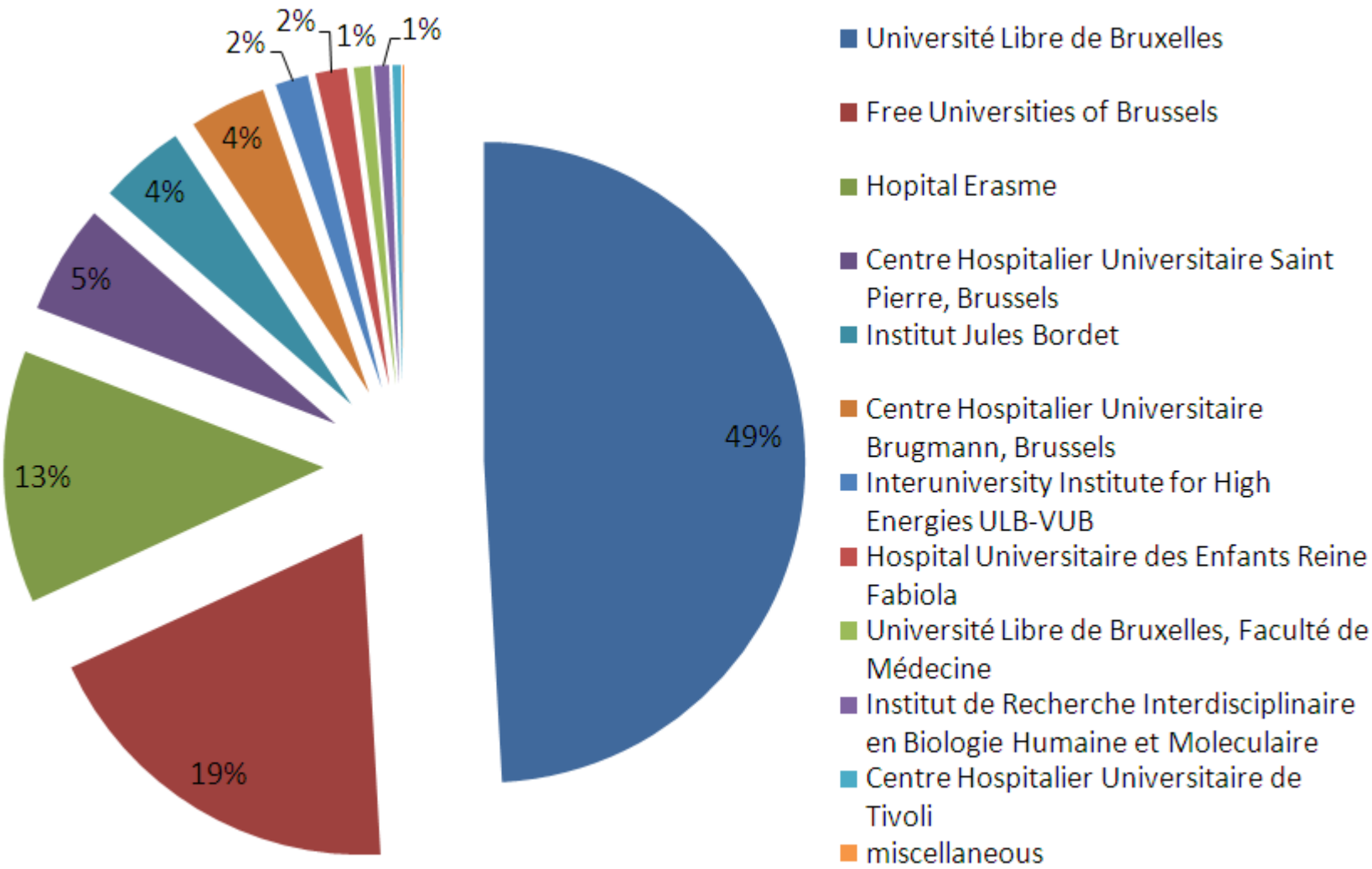
- DI-fusion: institutional repository launched in 2009
  - Deployment calendar following the research assessment process currently in progress
  - Mandatory for every researcher to be referenced
- Libraries provide support to the community by
  - Encoding references manually
  - Automatically importing metadata from external sources

→ **Partnership with Elsevier**

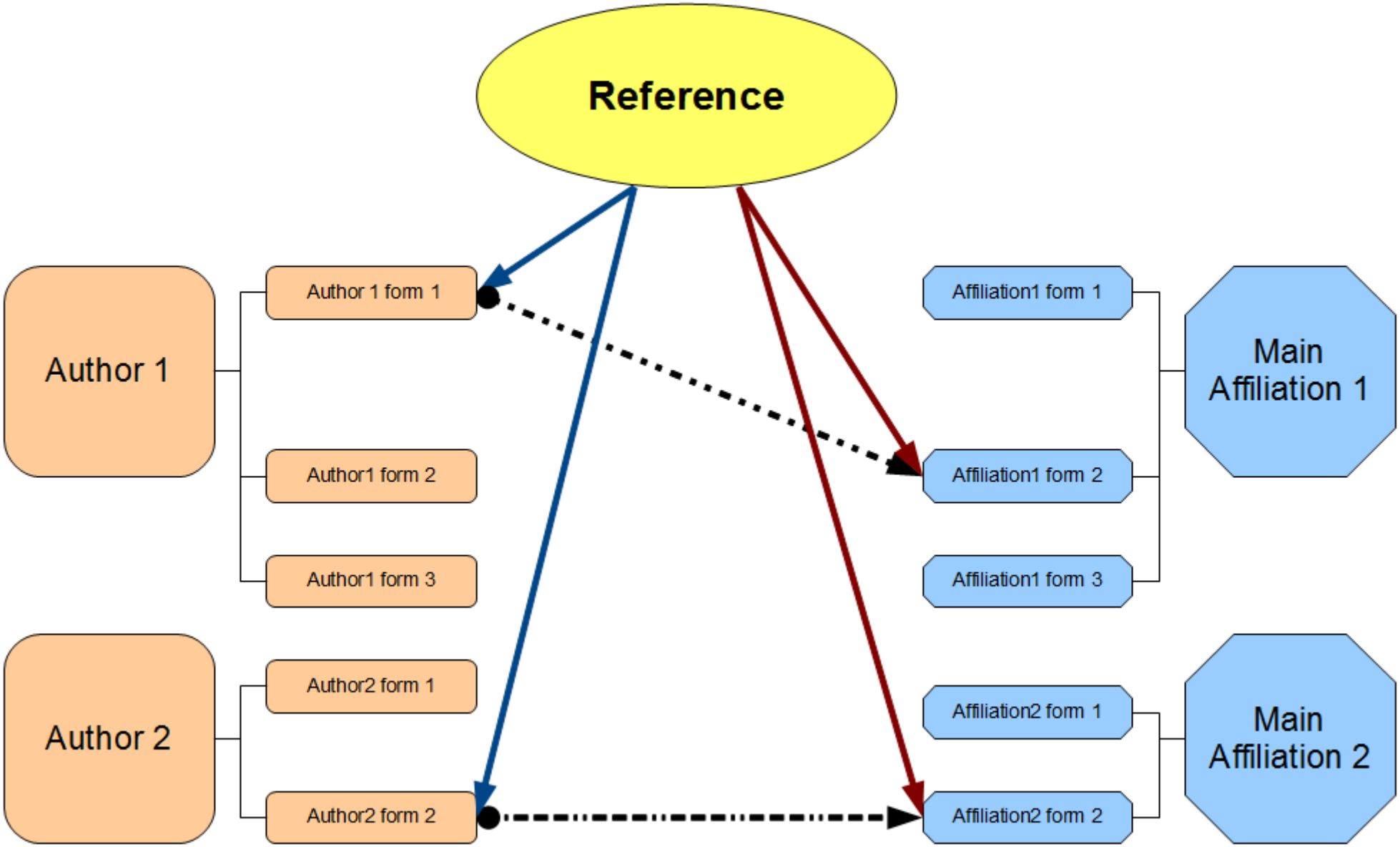
Context

- Partnership with Elsevier (Scopus)
  - 49,000 references of ULB scholars in XML
  - Metadata extraction based on author affiliation
  - **43** distinct IDs matching ULB (manually found)
- Several data quality issues
- Need for <u>methods</u> and <u>tools</u>

Main ULB-related affiliations



Scopus architecture



Method

1. Qualitative analysis

- First glimpse at the extent and diversity of issues affecting metadata quality
- Sample of scholars based on criteria hypothesized to be correlated to metadata quality issues:
  - Broad research field
  - Number of published references
  - Name complexity
  - Homonymy

Method

2. Quantitative analysis

- Large CSV files containing references extracted from Scopus (sample of 43 known affiliations)
- Three successive steps:

  a) Data profiling

  b) Correction and cleaning

  c) Enrichment and export

Tool: Google Refine

- Free app to enhance and enrich messy data
- Run locally through a browser interface
- Powerful and multifunctional
  - Detection of doubles
  - Facetting and filtering
  - Clustering near-duplicates
  - Reconciliation with knowledge bases
  - Templating for export

Creating a project

Google refine

A power tool for working with messy data.

Open a Project

Name	Last modified
phm collection utf8	today 9:20 AM
scopus FUB global	yesterday 5:05 PM
scopus FUB	yesterday 3:26 PM
scopus kiss	2 days ago
scopus malaisse 2	2 days ago
refs scopus	2 days ago
scopus malaisse	4 days ago
scopus keywords	4 days ago
scopus kw2	4 days ago
cat counts	a month ago
afmu	a month ago

Browse workspace directory

Create a New Project

or Import an Existing Project

Data file: Choose File scopus.csv

or data file URL:

Project name: scopus

Create Project

Advanced Options

Limit load to: rows (blank for all)

Ignore: 0 initial non-blank lines

Skip: 0 initial data rows

When parsing text files:

☑ Split into columns

Column separator:
(leave blank to auto-detect)

☑ Auto-detect value types
(numbers, dates, etc)

Header lines: 1
(use 0 if your data has no header)

☐ Ignore quotation marks

- About Google Refine
- Project Home Page
- Screencasts
- Help Documentation

Version 2.0 [r1836]

Affiliations

Cluster & Edit column "Affiliations"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

Method key collision ▾ Keying Function fingerprint ▾ **2443** clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
12	275	• Université Libre de Bruxelles, Belgium (137 rows) • Université Libre de Bruxelles, Bruxelles, Belgium (78 rows) • Universite Libre de Bruxelles, Bruxelles, Belgium (40 rows) • Universite Libre de Bruxelles, Belgium (4 rows) • Université libre de Bruxelles, Belgium (4 rows) • Université Libre de Bruxelles, Belgium. (3 rows) • Universitè Libre de Bruxelles, Belgium (2 rows) • Universitè Libre de Bruxelles, Bruxelles, Belgium (2 rows) • Université libre de Bruxelles, Bruxelles, Belgium (2 rows) • Universite Libre De Bruxelles, Bruxelles, Belgium (1 rows) • Université Libre De Bruxelles, Bruxelles, Belgium (1 rows) • Université Libre de, Bruxelles, Belgium (1 rows)	☐	Université Libre de Bruxelles, E
11	29	• Optique Nonlinéaire Théorique, Université Libre de Bruxelles, Campus Plaine, CP 231, B-1050 Bruxelles, Belgium (12 rows) • Optique Nonlinéaire Théorique, Université Libre de Bruxelles, Campus Plaine CP 231, B-1050 Bruxelles,	☐	Optique Nonlinéaire Théorique,

 # Choices in Cluster

2 — 12

 # Rows in Cluster

0 — 280

Average Length of Choices

0 — 230

Length Variance of Choices

0 — 11

Select All | Deselect All | **Merge Selected & Re-Cluster** | Merge Selected & Close | Close

Authors

Cluster & Edit column "Authors"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

Method key collision ▾ Keying Function fingerprint ▾ **551** clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
4	13	• Hoffmann K.H. (9 rows) • Hoffmann K.-H. (2 rows) • Hoffmann K.h. (1 rows) • K-H Hoffmann (1 rows)	☐	Hoffmann K.H.
4	22	• D'Hondt J. (18 rows) • D'hondt J. (2 rows) • DHondt J. (1 rows) • Dhondt J. (1 rows)	☐	D'Hondt J.
4	30	• van Pottelsberghe de la Potterie B. (18 rows) • Van Pottelsberghe De La Potterie B. (10 rows) • Pottelsberghe De La Potterie Van B. (1 rows) • Van Pottelsberghe de la Potterie B. (1 rows)	☐	van Pottelsberghe de la Potteri
4	13	• De Biseau J.-C. (6 rows) • De Biseau J.C. (4 rows) • de Biseau J.-C. (2 rows) • De Biseau Jc. (1 rows)	☐	De Biseau J.-C.

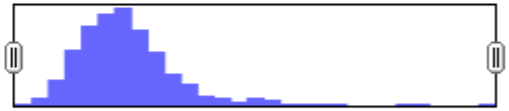
 # Choices in Cluster

2 — 4

 # Rows in Cluster

0 — 140

Average Length of Choices



7 — 35

Length Variance of Choices

0 — 5

Select All | Deselect All | **Merge Selected & Re-Cluster** | Merge Selected & Close | Close

Custom transformation

Custom text transform on column Affiliations

Expression — Language: Google Refine Expression Language (GREL)

```
if(value.contains("Bru"),"Université libre de Bruxelles,
Belgium",value)
```

No syntax error.

Preview | History | Help

7.	Laboratory of Experimental Surgery, Université Libre de Bruxelles, Brussels, Belgium	Université libre de Bruxelles, Belgium
8.	Institute of Human Nutrition, Columbia University, New York, NY, United States	Institute of Human Nutrition, Columbia University, New York, NY, United States
9.	Schools of Medicine, University of Manchester, United Kingdom	Schools of Medicine, University of Manchester, United Kingdom
10.	Life Sciences, University of Manchester, United Kingdom	Life Sciences, University of Manchester, United Kingdom

On error: (•) set to blank ( ) store error ( ) keep original

[ ] Re-transform up to 10 times until no change

OK | Cancel

Before/after transformation

Affiliations change

644 choices Sort by: name **count** Cluster

- Laboratory of Experimental Medicine, Brussels Free University, Brussels, Belgium 167
- Laboratory of Experimental Medicine, Brussels Free University, 808 Route de Lennik, B-1070 Brussels, Belgium 138
- Laboratory of Experimental Medicine, Brussels Free University, B-1000 Brussels, Belgium 55
- Laboratory of Experimental Medicine, Erasmus Medical School, Brussels Free University, 808 Route de Lennik, B-1070 Brussels, Belgium 39
- Lab. Exp. Med., Brussels Univ., Brussels, Belgium 24
- Laboratory of Experimental Hormonology, Brussels Free University, 808 Route de Lennik, B-1070 Brussels, Belgium 22
- Lab. Exp. Med., Brussels Univ. Sch. Med., Brussels, Belgium 21
- Lab. Exp. Med., Univ. Brussels, Belgium 19
- Laboratory of Experimental Medicine, Erasmus School of Medicine, Brussels Free University, 808 Route de Lennik, B-1070 Brussels, Belgium 19

Affiliations change

207 choices Sort by: name **count** Cluster

- Université libre de Bruxelles, Belgium 1282
- Leo Pharmaceutical Products, Ballerup, Denmark 9
- Biokimya Anabilim Dali, Cerrahpasa Tip Fakultesi, Istanbul Universitesi, Istanbul, Turkey 4
- Endocrinology and Diabetes Unit, Hospital Clinic, Barcelona, Spain 4
- Fund. Jiménez Díaz, Madrid, Spain 4
- Laboratoire de Nutrition Humaine, Clermont-Ferrand, France 4
- Ottawa Health Research Institute, University of Ottawa, Ottawa, Ont., Canada 4
- Department of Metabolism, Nutrition and Hormones, Fundación Jiménez Díaz, Avda. Reyes Católicos 2, 28040 Madrid, Spain 3
- Departments of Clinical Biochemistry and Medicine, University of Manchester, Manchester, United Kingdom 3
- Endocrinology Unit, Hospital Clinic, Barcelona, Spain 3
- Fundacion Jimenez Diaz, Madrid, Spain 3

Recursive clustering

Cluster & Edit column "Affiliations"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

Method key collision

Keying Function fingerprint
- fingerprint
- ngram-fingerprint
- double-metaphone

759 clusters found

Cluster Size	Row Count	Values in Cluster	Cell Value
11	21	• Department of Psychiatry, Erasme Hospital, Free University of Brussels, 808 route de Lennik, B-1070 Brussels, Belgium (5 rows) • Department of Psychiatry, Erasme Hospital, Free University of Brussels, 808 route de Lennik, B-1070, Brussels, Belgium (5 rows) • Department of Psychiatry, Erasme Hospital, Free University of Brussels, 808 Route de Lennik, B-1070 Brussels, Belgium (3 rows) • Department of Psychiatry, Erasme Hospital, Free University of Brussels, 808 Route de Lennik, B-1070, Brussels, Belgium (1 rows) • Department of Psychiatry, Erasme Hospital, Free University of Brussels, 808, Route de Lennik, B-1070 Brussels, Belgium (1 rows) • Department of Psychiatry, Erasme Hospital, Free University of Brussels, Route de Lennik 808, B-1070 Brussels, Belgium (1 rows) • Department of Psychiatry, Erasme Hospital, Free University of Brussels, Route de Lennik 808, B-1070, Brussels, Belgium (1 rows) • Department of Psychiatry, Free University of Brussels, Erasme Hospital, Route de Lennik 808, B-1070 Brussels, Belgium (1 rows) • Free University of Brussels, Department of	Department of Psychiatry, Eras

 # Choices in Cluster

2 — 11

 # Rows in Cluster

0 — 220

Average Length of Choices

20 — 180

Length Variance of Choices

0 — 19

Select All | Deselect All | **Merge Selected & Re-Cluster** | Merge Selected & Close | Close

ULB

Affiliation disambiguation

Add column based on column Affiliations

New column name: General affiliation

On error: set to blank / store error / keep original

Expression — Language: Google Refine Expression Language (GREL)

```
value.replace(/.*((ULB)|(Bruxelles)|([lL]ibre)|(Erasme)|
(Roosevelt)).*/,"ULB")
```

No syntax error.

Preview | History | Help

70.	Department of Intensive Care, Erasme Hospital, Free University of Brussels, Brussels, Belgium	ULB
72.	Hogeschool Universiteit Brussel, Erasmusgebouw, Stormstraat 2, B1000 Brussel, Belgium	Hogeschool Universiteit Brussel, Erasmusgebouw, Stormstraat 2, B1000 Brussel, Belgium
75.	Manual Therapy, Department of Rehabilitation and Physiotherapy, Free University of Brussels, Faculty of Medicine and	Manual Therapy, Department of Rehabilitation and Physiotherapy, Free University of Brussels, Faculty of Medicine and Pharmacology, Brussels, Belgium

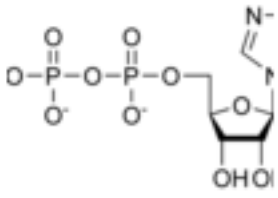
OK | Cancel

Reconciliation

	Affiliations	Authors with affiliations	Author Keywords
cord.url?eid=2-s2.0-=e7bbc9163ef392393f428c3356af17f7	Department of Metabolism,	Acitores, A., Department of Metabolism, Nutrition and	Glucose transport; Protein kinases

Match this Cell | Match All Identical Cells | Cancel

Adenosine triphosphate



Adenosine-5'-triphosphate (ATP) is a multifunctional nucleotide used in cells as a coenzyme. It is often called the "molecular unit of currency" of intracellular energy transfer. ATP transports...

more

Source: Freebase CC-BY

Related Links

- Wikipedia
- Italian Wikipedia
- Spanish Wikipedia
- German Wikipedia
- Japanese Wikipedia
- French Wikipedia
- Bulgarian Wikipedia
- Portugese Wikipedia
- Russian Wikipedia

Metaweb

Get one for any topic!

Adenosine triphosphate (4)
Insulin resistance (4)
Create new topic
or match

Dpt. Metabolismo, Nutrición y Hormonas, Fundación Jiménez Diaz, Avda Reyes Católicos 2, 28040 Madrid, Spain

Templating export

Templating Export

Prefix

```
{
  "rows" : [
```

Row Template

```
    {
      "Authors" : {{jsonize(cells["Authors"].value)}},
      "Title" : {{jsonize(cells["Title"].value)}},
      "Link" : {{jsonize(cells["Link"].value)}},
      "Affiliations" : {{jsonize(cells["Affiliations"
      "General affiliation" : {{jsonize(cells["Genera
      "Authors with affiliations" : {{jsonize(cells["
      "Author Keywords" : {{jsonize(cells["Author Key
      "Index Keywords" : {{jsonize(cells["Index Keywo
      "Correspondence Address" : {{jsonize(cells["Cor
      "Language of Original Document" : {{jsonize(cel
    }
```

Row Separator

```
,
```

Suffix

```
  ]
}
```


```
{
  "rows" : [
    {
      "Authors" : "Walravens N., Pauwels C.",
      "Title" : "From high hopes to high deficit and
      "Link" : "http://www.scopus.com/inward/record.
      "Affiliations" : "IBBT-SMIT, Free University c
      "General affiliation" : "VUB",
      "Authors with affiliations" : "Walravens, N.,
      "Author Keywords" : "Blu Ray; European policy;
      "Index Keywords" : "Blu-Ray; European policy;
      "Correspondence Address" : "Walravens, N.; IBE
      "Language of Original Document" : "English"
    },
    {
      "Authors" : "Cherchye L., Moesen W., Rogge N.,
      "Title" : "Constructing composite indicators w
      "Link" : "http://www.scopus.com/inward/record.
      "Affiliations" : "Katholieke Universiteit Leuv
      "General affiliation" : null,
      "Authors with affiliations" : "Cherchye, L., E
      "Author Keywords" : "Benefit of the doubt; Com
      "Index Keywords" : "Benefit of the doubt; Comp
      "Correspondence Address" : "Van Puyenbroeck, T
      "Language of Original Document" : "English"
    },
    {
      "Authors" : null,
      "Title" : null,
      "Link" : null,
      "Affiliations" : "CentER, Tilburg University,
      "General affiliation" : null,
      "Authors with affiliations" : null
```


Reset Template | Export | Cancel

Conclusions

- Operational framework for improving ULB scholars visibility
  - Complementary methods of analysis
  - Comprehensive assessment of data quality
  - Efficient tools to assist specialists
- Reproducible in similar contexts
- Full paper (to be published) contains practical recommendations for libraries/institutions