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 methods and tools
Main ULB-related affiliations

- Université Libre de Bruxelles: 49%
- Free Universities of Brussels: 13%
- Hopital Erasme: 19%
- Centre Hospitalier Universitaire Saint Pierre, Brussels: 4%
- Institut Jules Bordet: 4%
- Centre Hospitalier Universitaire Brugmann, Brussels: 2%
- Interuniversity Institute for High Energies ULB-VUB: 2%
- Hospital Universitaire des Enfants Reine Fabiola: 1%
- Université Libre de Bruxelles, Faculté de Médecine: 1%
- Institut de Recherche Interdisciplinaire en Biologie Humaine et Moleculaire: 1%
- Centre Hospitalier Universitaire de Tivoli: 1%
- Miscellaneous: 1%
Scopus architecture

Reference

Author 1
- Author 1 form 1
- Author 1 form 2
- Author 1 form 3

Author 2
- Author 2 form 1
- Author 2 form 2

Main Affiliation 1
- Affiliation 1 form 1
- Affiliation 1 form 2
- Affiliation 1 form 3

Main Affiliation 2
- Affiliation 2 form 1
- Affiliation 2 form 2
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

<table>
<thead>
<tr>
<th>Name</th>
<th>Last modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>phm collection utf8</td>
<td>today 9:20 AM</td>
</tr>
<tr>
<td>scopus FUB global</td>
<td>yesterday 5:05 PM</td>
</tr>
<tr>
<td>scopus FUB</td>
<td>yesterday 3:26 PM</td>
</tr>
<tr>
<td>scopus kiss</td>
<td>2 days ago</td>
</tr>
<tr>
<td>scopus malasse 2</td>
<td>2 days ago</td>
</tr>
<tr>
<td>refs scopus</td>
<td>2 days ago</td>
</tr>
<tr>
<td>scopus malasse</td>
<td>4 days ago</td>
</tr>
<tr>
<td>scopus keywords</td>
<td>4 days ago</td>
</tr>
<tr>
<td>scopus kw2</td>
<td>4 days ago</td>
</tr>
<tr>
<td>cat counts</td>
<td>a month ago</td>
</tr>
<tr>
<td>afmu</td>
<td>a month ago</td>
</tr>
</tbody>
</table>

Create a New 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)

When parsing text files:

- Split into columns

Ignore:

- initial non-blank lines

Skip:

- initial data rows

Header lines: 1

(leave blank to auto-detect)

Column separator: 

(leave blank to auto-detect)

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

Ignore quotation marks

Version 2.0 [r1836]

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

### Cluster Size: 12  
**Row Count**: 275

- Université Libre de Bruxelles, Belgium (137 rows)
- Université Libre de Bruxelles, Bruxelles, Belgium (78 rows)
- Université Libre de Bruxelles, Bruxelles, Belgium (40 rows)
- Université Libre de Bruxelles, Belgium (4 rows)
- Université libre de Bruxelles, Belgium (3 rows)
- Université Libre de Bruxelles, Bruxelles, Belgium (2 rows)
- Université Libre de Bruxelles, Bruxelles, Belgium (2 rows)
- Université libre de Bruxelles, Bruxelles, Belgium (2 rows)
- Université Libre De Bruxelles, Bruxelles, Belgium (1 rows)
- Université Libre De Bruxelles, Bruxelles, Belgium (1 rows)
- Université Libre de Bruxelles, Bruxelles, Belgium (1 rows)

**Merge?**  
**New Cell Value**: Université Libre de Bruxelles, Belgium

### Cluster Size: 11  
**Row Count**: 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, Belgium (1 rows)

**Merge Selected & Re-Cluster**  
**Merge Selected & Close**  
**Close**

---

**# Choices in Cluster**  
2 — 12

**# Rows In Cluster**  
0 — 280

**Average Length of Choices**  
0 — 230

**Length Variance of Choices**  
0 — 11
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

<table>
<thead>
<tr>
<th>Cluster Size</th>
<th>Row Count</th>
<th>Values in Cluster</th>
<th>Merge?</th>
<th>New Cell Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>13</td>
<td>Hoffmann K.H. (9 rows) Hoffmann K.H. (2 rows) Hoffmann K.h. (1 row) K-H Hoffmann (1 row)</td>
<td>□</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>22</td>
<td>D'Hondt J. (18 rows) D'hondt J. (2 rows) D'Hondt J. (1 row) Dhondt J. (1 row)</td>
<td>□</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>van Pottelsberghe de la Potterie B. (18 rows) Van Pottelsbergh De La Potterie B. (10 rows) Pottelsbergh De La Potterie Van B. (1 row) Van Pottelsberghe de la Potterie B. (1 row)</td>
<td>□</td>
<td></td>
</tr>
</tbody>
</table>

# Choices in Cluster

# Rows In Cluster

Average Length of Choices

Length Variance of Choices

Select All  Deselect All  Merge Selected & Re-Cluster  Merge Selected & Close  Close
Custom text transform on column Affiliations

<table>
<thead>
<tr>
<th>Expression</th>
<th>Language</th>
<th>CustomExpression</th>
<th>Preview</th>
<th>History</th>
<th>Help</th>
</tr>
</thead>
<tbody>
<tr>
<td>if(value.contains(&quot;Bru&quot;),&quot;Université libre de Bruxelles, Belgium&quot;,value)</td>
<td>Google Refine Expression Language (GREL)</td>
<td>No syntax error.</td>
<td>Université libre de Bruxelles, Belgium</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Affiliations</th>
<th>Universities/Institutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>7. Laboratory of Experimental Surgery, Université Libre de Bruxelles, Brussels, Belgium</td>
<td>Université libre de Bruxelles, Belgium</td>
</tr>
<tr>
<td>8. Institute of Human Nutrition, Columbia University, New York, NY, United States</td>
<td>Institute of Human Nutrition, Columbia University, New York, NY, United States</td>
</tr>
<tr>
<td>9. Schools of Medicine, University of Manchester, United Kingdom</td>
<td>Schools of Medicine, University of Manchester, United Kingdom</td>
</tr>
<tr>
<td>10. Life Sciences, University of Manchester, United Kingdom</td>
<td>Life Sciences, University of Manchester, United Kingdom</td>
</tr>
</tbody>
</table>

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

Re-transform up to 10 times until no change
Before/after transformation

<table>
<thead>
<tr>
<th>Affiliations</th>
<th>change</th>
</tr>
</thead>
<tbody>
<tr>
<td>644 choices</td>
<td></td>
</tr>
<tr>
<td>Sort by: name count</td>
<td></td>
</tr>
<tr>
<td>Cluster</td>
<td></td>
</tr>
<tr>
<td>Laboratory of Experimental Medicine, Brussels Free University, Brussels, Belgium</td>
<td>167</td>
</tr>
<tr>
<td>Laboratory of Experimental Medicine, Brussels Free University, 808 Route de Lennik, B-1070 Brussels, Belgium</td>
<td>138</td>
</tr>
<tr>
<td>Laboratory of Experimental Medicine, Brussels Free University, B-1000 Brussels, Belgium</td>
<td>55</td>
</tr>
<tr>
<td>Laboratory of Experimental Medicine, Erasmus Medical School, Brussels Free University, 808 Route de Lennik, B-1070 Brussels, Belgium</td>
<td>39</td>
</tr>
<tr>
<td>Laboratory of Experimental Hormonology, Brussels Free University, 808 Route de Lennik, B-1070 Brussels, Belgium</td>
<td>22</td>
</tr>
<tr>
<td>Laboratory of Experimental Medicine, Erasmus School of Medicine, Brussels Free University, 808 Route de Lennik, B-1070 Brussels, Belgium</td>
<td>19</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Affiliations</th>
<th>change</th>
</tr>
</thead>
<tbody>
<tr>
<td>207 choices</td>
<td></td>
</tr>
<tr>
<td>Sort by: name count</td>
<td></td>
</tr>
<tr>
<td>Cluster</td>
<td></td>
</tr>
<tr>
<td>Université libre de Bruxelles, Belgium</td>
<td>1282</td>
</tr>
<tr>
<td>Leo Pharmaceutical Products, Ballerup, Denmark</td>
<td>9</td>
</tr>
<tr>
<td>Biokimya Anabilim Dali, Cerrahpasa Tip Fakultesi, Istanbul Universitesi, Istanbul, Turkey</td>
<td>4</td>
</tr>
<tr>
<td>Endocrinology and Diabetes Unit, Hospital Clinic, Barcelona, Spain</td>
<td>4</td>
</tr>
<tr>
<td>Fund. Jiménez Díaz, Madrid, Spain</td>
<td>4</td>
</tr>
<tr>
<td>Laboratoire de Nutrition Humaine, Clermont-Ferrand, France</td>
<td>4</td>
</tr>
<tr>
<td>Ottawa Health Research Institute, University of Ottawa, Ottawa, Ont., Canada</td>
<td>4</td>
</tr>
<tr>
<td>Department of Metabolism, Nutrition and Hormones, Fundación Jiménez Díaz, Avda. Reyes Católicos 2, 28040 Madrid, Spain</td>
<td>3</td>
</tr>
<tr>
<td>Departments of Clinical Biochemistry and Medicine, University of Manchester, Manchester, United Kingdom</td>
<td>3</td>
</tr>
<tr>
<td>Endocrinology Unit, Hospital Clinic, Barcelona, Spain</td>
<td>3</td>
</tr>
<tr>
<td>Fundacion Jimenez Díaz, Madrid, Spain</td>
<td>3</td>
</tr>
</tbody>
</table>
Recursive clustering

Cluster & Edit column "Affiliations"

This feature helps you find groups of different cell values that might be alternative representations of the samething. 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 ...

<table>
<thead>
<tr>
<th>Method</th>
<th>Key Collision</th>
<th>Keying Function</th>
<th>Cell Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>key collision</td>
<td>fingerprint</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster Size</th>
<th>Row Count</th>
<th>Values in Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>21</td>
<td>- Department of Psychiatry, Erasme Hospital, Free University of Brussels, 808 route de Lennik, B-1070 Brussels, Belgium (3 rows)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Department of Psychiatry, Erasme Hospital, Free University of Brussels, 808 route de Lennik, B-1070, Brussels, Belgium (5 rows)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Department of Psychiatry, Erasme Hospital, Free University of Brussels, 808 route de Lennik, B-1070 Brussels, Belgium (3 rows)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Department of Psychiatry, Erasme Hospital, Free University of Brussels, 808 Route de Lennik, B-1070 Brussels, Belgium (1 rows)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Department of Psychiatry, Erasme Hospital, Free University of Brussels, 808, Route de Lennik, B-1070 Brussels, Belgium (1 rows)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Department of Psychiatry, Erasme Hospital, Free University of Brussels, Route de Lennik 808, B-1070 Brussels, Belgium (1 rows)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Department of Psychiatry, Erasme Hospital, Free University of Brussels, Route de Lennik 808, B-1070 Brussels, Belgium (1 rows)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Free University of Brussels, Department of Psychiatry, Erasme Hospital, Free University of Brussels, Route de Lennik 808, B-1070 Brussels, Belgium (1 rows)</td>
</tr>
</tbody>
</table>

759 clusters found

- # Choices in Cluster
- # Rows in Cluster
- Average Length of Choices
- Length Variance of Choices
Affiliation disambiguation

![Image](https://example.com/image.png)

**Add column based on column Affiliations**

- **New column name**: General affiliation
- **On error**:
  - [ ] set to blank
  - [ ] store error
  - [ ] keep original
- **Expression**:
  ```grel
  value.replace(/.*((ULB)|Bruxelles)|([IL]ibre)|(Erasme)|Roosevelt).*="/ULB""
  ```

**Preview**

<table>
<thead>
<tr>
<th>ID</th>
<th>Affiliation</th>
<th>University</th>
</tr>
</thead>
<tbody>
<tr>
<td>70</td>
<td>Department of Intensive Care, Erasme Hospital, Free University of Brussels, Brussels, Belgium</td>
<td>ULB</td>
</tr>
<tr>
<td>72</td>
<td>Hogeschool Universiteit Brussel, Erasmusgebouw, Stormstraat 2, B1000 Brussel, Belgium</td>
<td>Hogeschool Universiteit Brussel, Erasmusgebouw, Stormstraat 2, B1000 Brussel, Belgium</td>
</tr>
<tr>
<td>75</td>
<td>Manual Therapy, Department of Rehabilitation and Physiotherapy, Free University of Brussels, Faculty of Medicine and Pharmacology, Brussels, Belgium</td>
<td>Manual Therapy, Department of Rehabilitation and Physiotherapy, Free University of Brussels, Faculty of Medicine and Pharmacology, Brussels, Belgium</td>
</tr>
</tbody>
</table>
Reconciliation

Adenosine 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...

Related Links

- Wikipedia
- Italian Wikipedia
- Spanish Wikipedia
- German Wikipedia
- Japanese Wikipedia
- French Wikipedia
- Bulgarian Wikipedia
- Portuguese Wikipedia
- Russian Wikipedia
Templating export

**Prefix**

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

**Row Template**

```
{
  "Authors" : {{jsonize(cells["Authors").value}}
```

**Row Separator**

```
,
```

**Suffix**

```
] }
```
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