

Leading The Way



# Scopus data and Affiliation Handling in Scopus

M'hamed el Aisati, Elsevier  
Tunis, 06 October 2016

# Agenda

1. Scopus
  - Coverage
  - Data model (high level)
2. Affiliation Handling in Scopus
  - Affiliation Profiler
  - Affiliation Profile Quality Measures
  - Feedback and manual corrections
  - Affiliation Profile Correction Types

# Agenda

1. Scopus
  - Coverage
  - Data model (high level)
2. Affiliation Handling in Scopus
  - Affiliation Profiler
  - Affiliation Profile Quality Measures
  - Feedback and manual corrections
  - Affiliation Profile Correction Types

# Scopus includes content from more than 5,000 publishers and 105 different countries

64M records from 23K serials, 90K conferences and 120K books

- Updated daily
- Records back to 1823
- “Articles in Press” from > 3,750 titles
- >50 different languages covered
- 3,715 active Gold Open Access journals indexed
- **Deep citation linking** for all articles 1970-present
- Authoritative **Author & Affiliation Profiles** for all records 1823-forward
- Additional **enhanced metadata**, ex. Medline & other index terms, Funding Acknowledgements, etc.

## JOURNALS

Physical  
Sciences

7,443

Health  
Sciences

6,795

Social  
Sciences

8,086

Life  
Sciences

4,492

21,568 peer-reviewed journals

361 trade journals

- Full metadata, abstracts and cited references (ref's post-1995 only)
- Funding data from acknowledgements
- Citations back to 1970

## CONFERENCES

90K conference events

7.3M conference papers

Mainly Engineering and Computer Sciences

## BOOKS

531 book series

30K Volumes /

1.2M items

119,882 stand-alone books

974K items

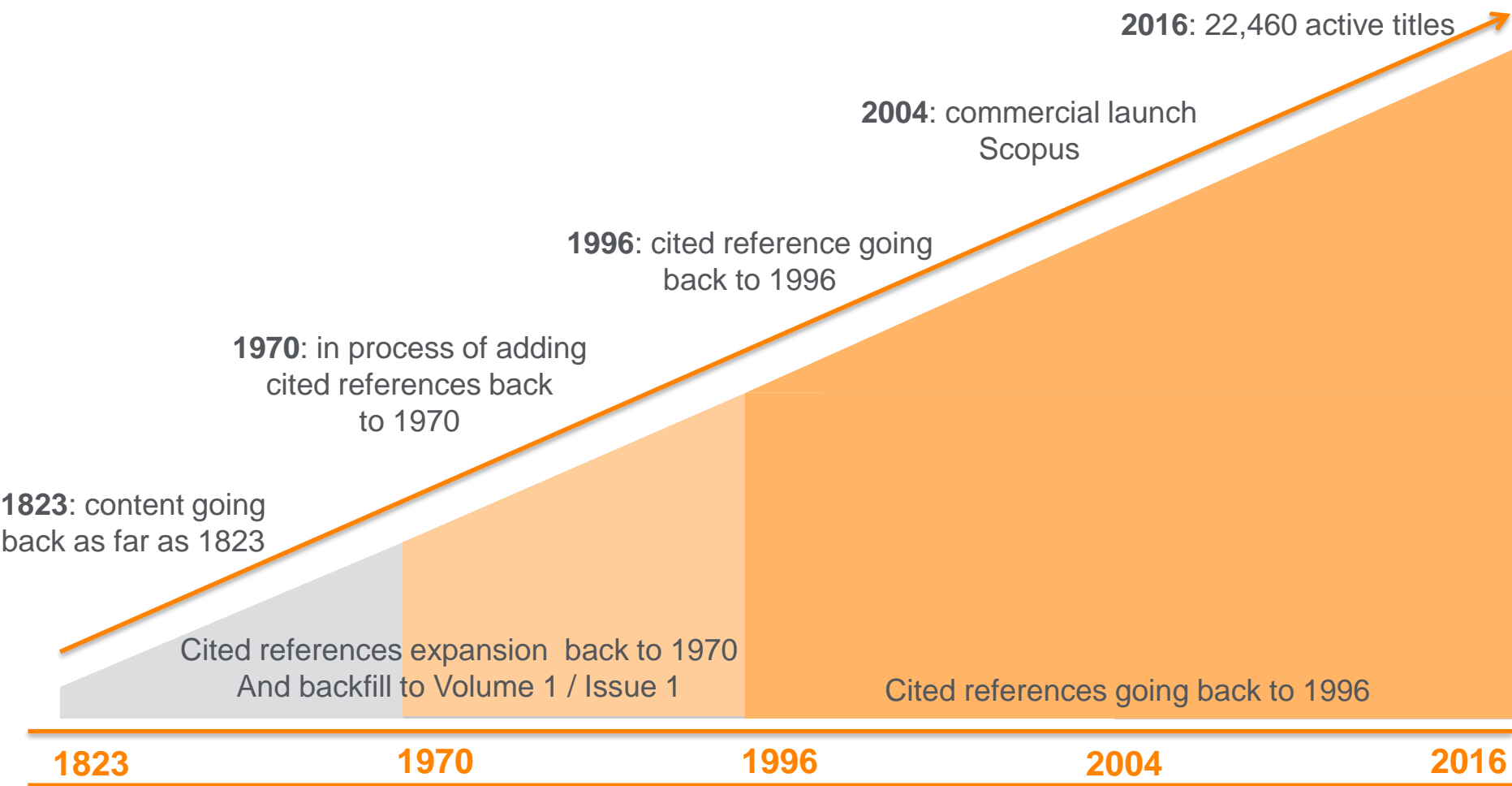
Focus on Social Sciences and A&H

## PATENTS\*

27M patents

From 5 major patent offices  
- WIPO  
- EPO  
- USPTO  
- JPO  
- UK IPO

# Scopus content has evolved over the last 12 years



## Comparison with nearest peer

### Scopus

~22K titles

>5,000 publishers

Updated daily

**Scopus**  
22,245

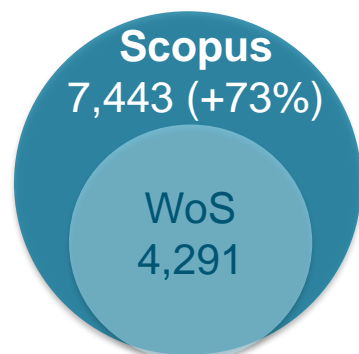
**Web of Science**  
12,140

### WEB OF SCIENCE™

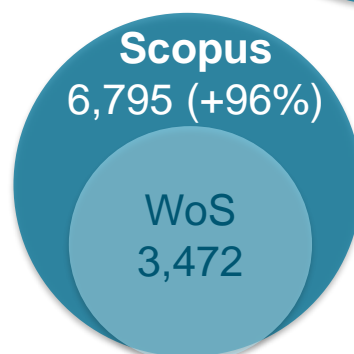
~12K titles (Core Collection)

3,300 publishers

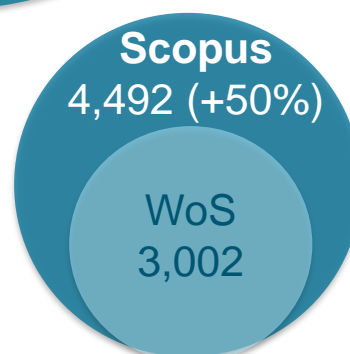
Updated weekly



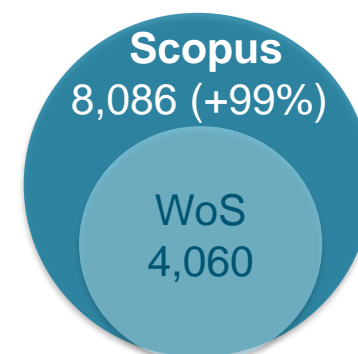
Physical Sciences



Health Sciences

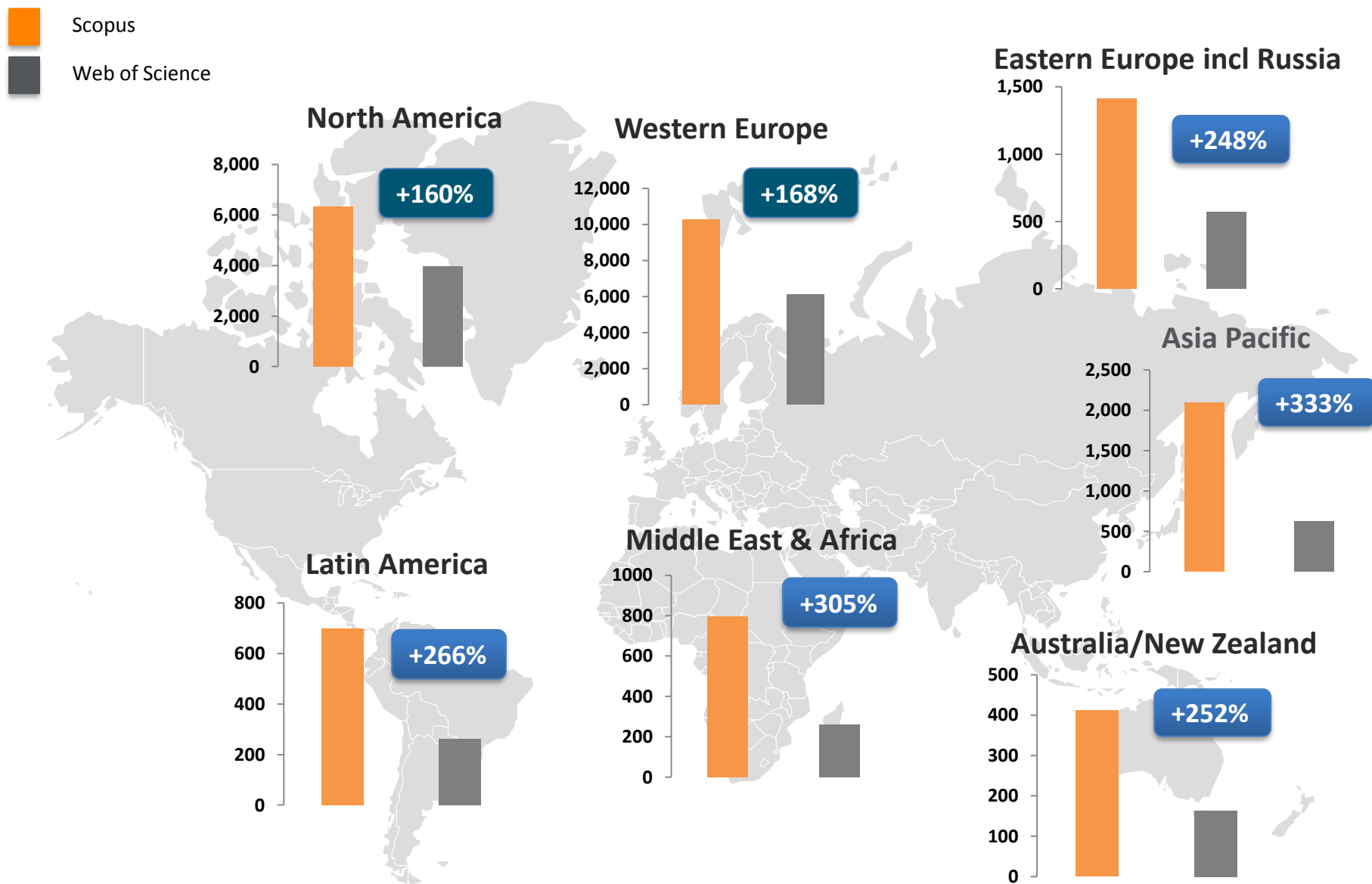


Life Sciences



Social Sciences

# Scopus: Breadth of Coverage Across Geographies



## Let's go back to basics: the Scopus data model

The **Scopus data model** is designed around the notion that **articles** are written by **authors** that are **affiliated** with **institutions**. Visually and rather simplistically, this relational model is represented below.

**What is the value of this structured data?** This relational data model means that Scopus can tell you **who is doing what** in global literature and **where they are doing it** with **higher accuracy** than anyone else

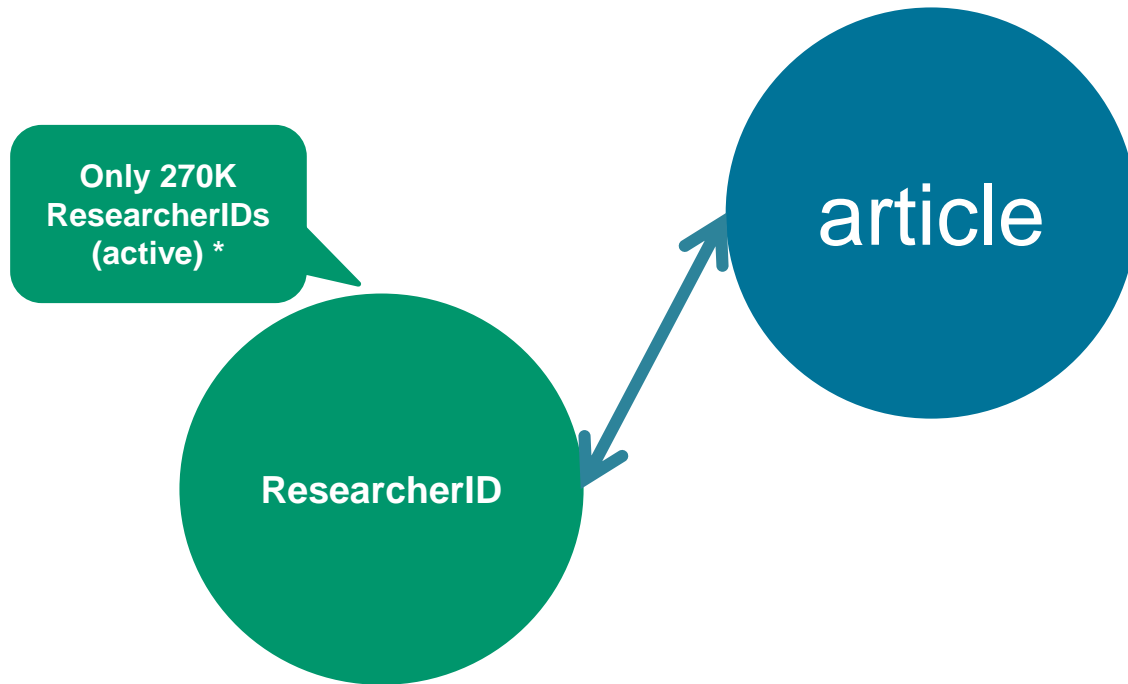


Scopus Data Model Simplified



# How does this compare?

HERE IS WHAT WOS DOES



\* Source: <http://wokinfo.com/researcherid/>

# Agenda

1. Scopus
  - Coverage
  - Data model (high level)
2. Affiliation Handling in Scopus
  - Affiliation Profiler
  - Affiliation Profile Quality Measures
  - Feedback and manual corrections
  - Affiliation Profile Correction Types

## Affiliation Profiling

Objective:

1. When processing an incoming article in the Scopus repository, assign unique identifier to an organization (e.g. university, research institute, government organization, hospital,...) that performed or sponsored research in order to effectively cluster all articles to its respective organization.
2. Cluster multiple names and addresses of organizations together.

## Affiliation Profiler Algorithm Steps

1. Creation of affiliation string from incoming article
2. Parsing and tagging of affiliation string
3. Affiliation Lookup in OrgDB\*
4. Handling of unmatched affiliations

\*OrgDB is a database which contains preferred name, name variants, address, hierarchy information for over 70,000 organizations worldwide

- Compiled from extensive manual effort
- Continually updated to add new organizations

## Step 1 – Creation of Affiliation String from incoming Article record

### Author & Affiliation info in incoming article:

```
<author-group><author seq="1" >PourabdoIhossein Fereshteh </author><author seq="2">Mozafari Sabah</author><author seq="3" Javan Mohamad </author><author seq="4"> Mirnajafi-zadeh Javad, mjavaan@modares.ac.ir </author><affiliation>Dept. Physiology, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran</affiliation></author-group><author-group><author seq="5" >Ahmadiani Abolhassan</author><affiliation country="irn" ><organization>Dept. Physiology</organization><organization>Faculty of Medical Sciences</organization><organization>Tarbiat Modares University</organization><city-group>Tehran</city-group></affiliation></author-group>
```



### Extracted Affiliation String:

```
<affiliation country="irn" ><organization>Dept. Physiology</organization><organization>Faculty of Medical Sciences</organization><organization>Tarbiat Modares University</organization><city-group>Tehran</city-group></affiliation>
```

## Step 2 – Parsing & Tagging Affiliation String

Incoming  
Affiliation  
String

- `<affiliation country="irn" ><organization>Dept. Physiology</organization><organization>Faculty of Medical Sciences</organization><organization>Tarbiat Modares University</organization><city-group>Tehran</city-group></affiliation>`

Affiliation  
Parser

- Identifies main organization and its sub from affiliation string for look up of it against OrgDB in next step

Formatted  
Affiliation  
String

- `<affiliation><sub-org>Dept. Physiology</sub-org><sub-org> Faculty of Medical Sciences</sub-org> <sup-org>Tarbiat Modares University<sup-org><city> Tehran</city></affiliation>`

## Organization Database (OrgDB)

- Core component of the Scopus Affiliation profiler
- OrgDB contains preferred name, name variants, address, hierarchy information for over **70,000 organizations worldwide**
  - Compiled from extensive manual effort
  - Continually updated to add new organizations

## Step 3 – Affiliation Lookup in OrgDB

### Formatted Affiliation String:

```
<affiliation><sub-org>Dept. Physiology</sub-org><sub-org> Faculty of Medical Sciences</sub-org> <sup-org>Tarbiat Modares University<sup-org><city>tehran</city><country>Iran</country></affiliation>
```

### Exported Affiliation ID in Scopus record:

```
<affiliation country="irn" afid="60032053" dptid="104626835"><organization>Dept. Physiology</organization><organization>Faculty of Medical Sciences</organization><organization>Tarbiat Modares University</organization><city-group>Tehran</city-group></affiliation>
```



### OrgDB Lookup:

1. Matches affiliation info in incoming string to find ID from OrgDB based on match of name, address, city, country



yes



no



See Step 4 on next slide



## Step 4 – Handling Unmatched Affiliations

Affiliation info in formatted string not in OrgDB

e.g. <affiliation> Kiev Lenin Order Kiev Engineering Institute, Russia </affiliation>



1. Cluster affiliation info of incoming string with other un-mapped affiliations
2. Assign un-mapped ID for internal use and stamp it in scopus article
3. Store clustered un-mapped affiliations



e.g. <affiliation country="rus" afid="100343523"><organization>Kiev Lenin Order Polytechnic Institute</organization></affiliation>

There is no affiliation profile for this. So, this affiliation can't be found in affiliation search. Only documents for it can be retrieved via affiliation name or ID search.



Periodical manual review of un-mapped affiliation info and its addition to OrgDB



Create affiliation profile and re-Export all affected articles with affiliation identifier info when un-mapped affiliation ID becomes mapped in periodical review

e.g. <affiliation country="rus" afid="60345673"><organization>Kiev Lenin Order Polytechnic Institute</organization></affiliation>

## Mapped IP vs. Unmapped IP

- The Affiliation Profiler software clusters affiliations of authors into Institution Profiles (also referred to as Affiliation Profiles).
- The majority of the clustering is done by the software mapping an incoming affiliation to an OrgDB entry. The software uses data contained in the OrgDB to do this mapping with very high accuracy. All affiliations mapped to the same OrgDB entry are then implicitly clustered.
- Institution Profiles composed of affiliations clustered via the mapping to OrgDB entries are called Mapped IPs. **The IDs of Mapped IPs are between 60,000,000 and 99,999,999.**
- Those affiliations that could not be mapped to OrgDB are then automatically clustered by the software. Each such cluster also corresponds to an Institution Profile and are called Unmapped IPs. **Unmapped IPs have IDs greater than 100,000,000.**
- **Over 86% of affiliations in AW are currently mapped to an OrgDB entry.**
- Mapped IP have very high quality as they are based on manually curated data in the OrgDB. In addition, Mapped IPs also have the benefit of the rich OrgDB content (e.g. hierarchy information, relationship type, org-type, clean names etc.)

## Mapped IP Examples: University of Pennsylvania

- Following are example affiliations that were mapped to the top level OrgDB entry **University of Pennsylvania** (60006297)
- Affiliations with only one organizational reference
  - **University of Pennsylvania**, 901 Stellar-Chance Laboratories, 422 Curie Boulevard, Philadelphia, PA 19104-6100
  - **Univ of Pennsylvania**, Philadelphia
- Affiliations that include non-stand-alone sub-orgs (i.e. sub-orgs that never occur by themselves in an affiliation without the parent being also present)
  - **Univ of Pennsylvania**, Dep of Comput, and Inf Sci, Philadelphia, PA
  - Dept. of Pathol. and Lab. Medicine, **Univ. of Pennsylvania**, Philadelphia, PA 19104
- Affiliations that include sub-orgs in OrgDB. Even if the software fails to correctly identify the sub-org, it can still correctly map to the main University of Pennsylvania
  - **Univ. of Pennsylvania**, VAMC, Philadelphia, PA
  - Primary Care Residency Program, Sch. Med., **Univ. Pennsylvania**, Philadelphia, Pa.
- Affiliations that include sub-orgs not in OrgDB. The software still succeeds in correctly mapping to the main University of Pennsylvania
  - Towne School of Civil and Mechanical Engineering, **University of Pennsylvania**, Philadelphia, PA 19104

# Mapped IP Examples: University of Pennsylvania School of Medicine

- Following are example affiliations that were mapped to **University of Pennsylvania School of Medicine** (IP 60003711). This is a sub-org **University of Pennsylvania** (60006297) whose main is set to true
- Affiliations with only one organizational reference
  - Univ. of Pennsylvania Sch. of Med., 421 Curie Boulevard, Philadelphia, PA 19104
  - University of Pennsylvania School of Medicine Philadelphia
- Affiliations that include non-stand-alone sub-orgs (i.e. sub-orgs that never occur by themselves in an affiliation without the parent being also present)
  - Dept. Orthop. Surg., **Univ. Pennsylvania Sch. Med.**, Philadelphia, Pa.
  - Department of Physiology, A700 Richards Bldg., **Univ. of PA School of Medicine**, Philadelphia, PA 19104-6085
- Affiliations strings contain sub-orgs the software was not able to match to an existing OrgDB entry but is able to match it to the correct prominent organization.
  - Signal Transduction Program, **Abramson Family Cancer Research Institute**, **University of Pennsylvania School of Medicine**, Philadelphia, PA 19104
  - **Monell Chem. Senses Cent.**, **Univ. Pennsylvania Sch. Med.**, Philadelphia, Pa. 19104

## Unmapped IPs

- When the software cannot find a match between an affiliation and an OrgDB entry, it automatically clusters the affiliation with others it determines are similar to create the Unmapped IPs.
- Here are 3 examples of Unmapped IPs each with **Wharton** in the name of the IP. The name of the IP is inferred by voting over all the affiliations that constitute the cluster.
  - **Wharton School** (100738323) is an Unmapped IP clustering 18 affiliations; 4 are reproduced below. Interestingly 17 of the affiliations have address information that imply they belong to **University of Pennsylvania** but none explicitly has the prominent org in the affiliation itself.
    - HERMES Laboratory for Financial Modeling and Simulation, Decision Sciences Department, The Wharton School, Philadelphia, PA 19104
    - Wharton School, Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340
    - School of Medicine, Wharton School, Philadelphia, USA
    - Department of Operations and Information Management, Wharton School, Philadelphia, PA 19104
  - **Wharton School of Business** (IP 100413267). None of the affiliation strings in this unmapped IP have address/location information, i.e. no address, city, state, or country.
    - Wharton School of Business
  - **The Wharton School** (101995705). Only 2 affiliations in this cluster have address/location information - San Francisco as the city. A web search confirms that there is an extension of **University of Pennsylvania Wharton School** in San Francisco. Though OrgDB currently does not have an entry for this, it could be added.
    - Department of Health Care Systems, Wharton School, San Francisco, CA

## Affiliation Profile Quality Measures

- **Precision**

- Precision is defined as average percentage (%) of articles that belong to same affiliation profile

- **Recall**

- Recall is defined as average percentage (%) of organization's publications that are in the organization's largest single profile

Current Levels for entire dataset:

**Precision** – 99 (+/- 1) %, **Recall** – 95 (+/- 1) %

Measurement is conducted twice a year.

# Feedback and manual correction is needed for more accurate profiles (1)

- Variations in incoming data makes it impossible to profile with 100% accuracy



1

**Ceské vysoké učení technické v Praze**

Czech Technical University  
Czech Technical University in Prague  
[Find unmatched affiliations](#)

9051

Prague




Czech Republic

Many articles from several authors contained institute names as one of below in article incoming. Such variations make it impossible to make one single profile. Based on incoming, we link articles from all these variations to main profile above.


- Dept. of Technical Mathematics CTU Prague
- FEE CTU Prague
- CVUT Praha
- IEAP CTU
- Faculty of Mechanical Engineering CTU
- CTU-Ericsson-Vodafone Research and Development Centre (RDC)
- České Vysoké Učení Technické

## Feedback and manual correction is needed for more accurate profiles (2)



- Even after using all possible disambiguating criteria, there are cases where multiple affiliations match on all criteria

<input type="checkbox"/>	<b>King's College London</b> 1 King's College London King's College <a href="#">Find unmatched affiliations</a>	 34770	London	United Kingdom
<input type="checkbox"/>	<b>King's College Institute of Psychiatry</b> 2 Institute of Psychiatry <a href="#">Find unmatched affiliations</a>	 17267	London	United Kingdom
<input type="checkbox"/>	<b>King's College School of Medicine and Dentistry</b> 3 <a href="#">Find unmatched affiliations</a>	 12575	London	United Kingdom

Separate profiles as names are different but customer would like to have it all consolidated together into one.

<input type="checkbox"/>	<b>University of Dayton Research Institute</b> 4 University of Dayton Research Institute Univ of Dayton Research Inst <a href="#">Find unmatched affiliations</a>	 1743	Dayton	United States
<input type="checkbox"/>	<b>University of Dayton Research Institute, Higley</b> 5 University of Dayton Research Institute <a href="#">Find unmatched affiliations</a>	 21	Higley	United States


These two could be one profile as one in Higley could be just a branch of one in Dayton or two separate one. Also, when articles don't contain city info – it becomes difficult to separate

<input type="checkbox"/>	<b>Coca-Cola</b> 24 Coca-Cola Company The Coca-Cola Company <a href="#">Find unmatched affiliations</a>	 177	Atlanta	United States
<input type="checkbox"/>	<b>Coca-Cola Enterprises</b> 25 Coca-Cola Enterprises Inc Coca-Cola Enterprises, Inc <a href="#">Find unmatched affiliations</a>	 6	Atlanta	United States


Separate profiles as names are different but potentially one.





# Implementing organizations' hierarchy: Search

Affiliation "pennsylvania"  Edit | SOLR Req/Res

44 affiliation results [About Scopus Affiliation Identifier](#)

Sort on: Document Count | Affiliation (A-Z) 

☐ All  Show documents |  Give feedback

Refine

Limit to

Exclude

City

☐ Philadelphia (14)

☐ Harrisburg (3)

☐ York (3)

☐ Pittsburgh (2)

☐ Abington (1)

Country/Territory

☐ United States (44)

Limit to

Exclude

[Export refine](#)

<input type="checkbox"/>	University of Pennsylvania 1 University of Pennsylvania Doc-XML   SOLR-JSON	147205	Philadelphia	United States
<input type="checkbox"/>	Pennsylvania State University 2 The Pennsylvania State University Pennsylvania State University Doc-XML   SOLR-JSON	141448	State College	United States
<input type="checkbox"/>	University of Pennsylvania, School of Medicine 3 Univ. of Pennsylvania Sch. of Med. University of Pennsylvania School of Medicine Doc-XML   SOLR-JSON	33430	Philadelphia	United States
<input type="checkbox"/>	University of Pennsylvania, Health System 4 Hosp. of the Univ. of Pennsylvania Univ. of Pennsylvania Medical Center Doc-XML   SOLR-JSON	26058	Philadelphia	United States
<input type="checkbox"/>	Penn State College of Medicine 5 Pennsylvania State University Pennsylvania State University College of Medicine Doc-XML   SOLR-JSON	18623	Hershey	United States
<input type="checkbox"/>	University of Pennsylvania, School of Veterinary Medicine 6 University of Pennsylvania Doc-XML   SOLR-JSON	7124	Kennett Square	United States
<input type="checkbox"/>	University of Pennsylvania, Wharton School 7 University of Pennsylvania Doc-XML   SOLR-JSON	6213	Philadelphia	United States
<input type="checkbox"/>	Indiana University of Pennsylvania 8 Indiana U Indiana University of Pennsylvania Doc-XML   SOLR-JSON	2476	Indiana	United States

# Implementing organizations' hierarchy: Display

Scopus

[Search](#) [Sources](#) [Alerts](#) [Lists](#) [Help](#) [M'hamed El Aisati](#)

## Affiliation details (University of Pennsylvania)

[Back to results](#) | 1 of 44 Next >

University of Pennsylvania  
3451 Walnut Street, Philadelphia  
PA, United States  
Affiliation ID: 60006297

[About Scopus Affiliation Identifier](#) | [View potential affiliation matches](#)  
Other name formats: University of Pennsylvania

Documents: 147,205  
Authors: 36,974

[Export](#) | [Print](#) | [E-mail](#)

[Follow this affiliation](#) Receive emails when new documents are available in Scopus.

[Set document feed](#)

[Give feedback about this affiliation](#)

[Documents by subject area](#)

☒ Hide Organizational Hierarchy: 12 of 19 [Expand full hierarchy](#)

**⚠** An affiliation refers to an organization or institution where an author conducts his/her primary research, unrelated to the source of the funding. Parent and sub-affiliations do not include external entities and do not pertain to sources of research funding.

<div> <input checked="" type="checkbox"/> Group with affiliation         </div> <div> <input checked="" type="checkbox"/> Page         </div> <div> <input type="checkbox"/> All         </div>		
Affiliations		
		Documents
	City	
→ University of Pennsylvania University of Pennsylvania		147205
<input type="checkbox"/> — University of Pennsylvania, Annenberg School for Communications University of Pennsylvania	Philadelphia	470
<input type="checkbox"/> — University of Pennsylvania, Law School University of Pennsylvania Law School	Philadelphia	449
<input type="checkbox"/> — <b>C</b> University of Pennsylvania, Health System Hosp. of the Univ. of Pennsylvania Univ. of Pennsylvania Medical Center Hospital of the University of Pennsylvania	Philadelphia	26058
<input type="checkbox"/> — <b>C</b> University of Pennsylvania, School of Medicine Univ. of Pennsylvania Sch. of Med. University of Pennsylvania School of Medicine	Philadelphia	33430
<input type="checkbox"/> — University of Pennsylvania, School of Nursing University of Pennsylvania School of Nursing	Philadelphia	1030
<input type="checkbox"/> — University of Pennsylvania, School of Engineering and Applied Science University of Pennsylvania School of Engineering and Applied Science	Philadelphia	32
<input type="checkbox"/> — University of Pennsylvania, School of Design University of Pennsylvania School of Design	Philadelphia	17
<input type="checkbox"/> — University of Pennsylvania, Graduate School of Education University of Pennsylvania Graduate School of Education	Philadelphia	28
<input type="checkbox"/> — University of Pennsylvania, School of Dental Medicine Univ. Pennsylvania University of Pennsylvania School of Dental Medicine	Philadelphia	746
<input type="checkbox"/> — University of Pennsylvania, School of Veterinary Medicine University of Pennsylvania School of Veterinary Medicine	Kennett Square	7124

## Scopus affiliation profiles vs. SciVal institution profiles

- By definition a Scopus affiliation profile = a SciVal institution profile
- However, there are cases where these are different
  1. Multiple separate Scopus profiles grouped as one institution profile in SciVal
  2. Different structure of an organization in Scopus vs. SciVal (based on client feedback)
  3. Changes applied to a profile in Scopus is not instantly reflected in SciVal (manual corrections)
- Process in place
  1. Change in Scopus
  2. Mapping adjusted in SciVal

## Affiliation Profiling Corrections

- **Types of corrections processed based on authoritative approval from an institute and news evidence**
  - Merges (combining two or more affiliation profiles together and updating associated Scopus articles as well)
  - Splits (separating articles from one affiliation profile to another affiliation profile (existing or new one))
  - Overrides (changing name, address, email of affiliation profile)
  - Hierarchy creation (establishing parent-child relationship between multiple affiliation profiles of same organization based on incoming)

# Thank you