Using VIVO, Scopus, and PubMed to disambiguate Weill Cornell authors

Paul Albert
paa2013@med.cornell.edu
Weill Cornell Medical College
Welcome to the VIVO for the CTSC

VIVO is a research-focused discovery tool that enables collaboration among scientists across all disciplines. VIVO contains information about researchers associated with the Clinical and Translational Science Center.

Browse or search for people, departments, courses, grants, and publications.

People (7,853)
- Faculty Member
- Faculty Member Emeritus
- Librarian
- Non-Academic
- Non-Faculty Academic
- Postdoc
- Professor Emeritus

Activities (1,612)
Courses (18)
Organizations (3,119)
Research (69,075)
Locations (396)
Laurie H. Glimcher is the Stephen and Suzanne Weiss Dean of Weill Cornell Medical College in New York, New York, where she is also Professor of Medicine. In addition, Dr. Glimcher is Provost for Medical Affairs of Cornell University. Previously, she was the Irene Heinz Given Professor of Immunology at the Harvard School of Public Health, where she was director of the Division of Biological Sciences, and Professor of Medicine at Harvard Medical School, where she headed one of the top immunology programs in the world. She also served as Senior Physician and Rheumatologist at the Brigham and Women’s Hospital.

Dr. Glimcher received her postdoctoral training at Harvard and in the Laboratory of Immunology at the National Institute of Allergy and Infectious Diseases in Bethesda, Maryland, and is board certified in Internal Medicine and Rheumatology. She received her BA degree from Radcliffe College and her MD from Harvard Medical School. As an immunologist, her primary research interests are the biochemical and genetic approaches that elucidate the molecular pathways that regulate CD4 T helper cell development and activation. The complex regulatory pathways governing T helper cell responses are critical for both the development of protective immunity and for the pathophysiologic immune responses underlying autoimmune, infectious and malignant diseases. Dr. Glimcher’s research laboratory has studied the transcriptional pathways...
that control this important immune checkpoint, leading to many discoveries, including the T-bet and XBP-1 transcription factors, which regulate a variety of immune functions. Most recently, her laboratory has identified new proteins that control osteoblast and osteoclast commitment and activation in skeletal biology with significant implications for diseases of bone, including osteoporosis, osteoarthritis and cancer metastasis to bone.

**PUBLICATIONS**

**Publications authored**


- Times cited: 34


- Times cited: 54


- Times cited: 88


- Times cited: 122

Original approach for managing faculty publications: rely on researchers or their proxies to manually enter publications.
Does this work?


17. HOGAN, E.L., CHIEN, J.L., AND DASGUPTA, S. GLYCOSPHYNGOLIPIDS CHICKEN SKELETAL MUSCLE IN EARLY DEVELOPMENT AND GENETIC DYSTROPHY. GANGLIOSIDE STRUCTURE, FUNCTION AND BIOMEDICAL POTENTIAL (ROBERT W. LEDEN, ROBERT K. YU, MAURICE M. RAPPORT AND


Um, what??
See those obnoxious diacritics? You’re welcome.
Researchers’ response to email requesting copy of CV

100% Responded

Did not respond
Why don’t our researchers care?

Failing to rigorously maintain an accurate list of publications is a rational choice. Time spent on maintaining publications bears a perceived, but more often real, opportunity cost.
Revised approach for managing faculty publications: use data from Scopus and PubMed to maintain profiles for them
Our publication ingest workflow

1. Librarian formulates queries. Stores in Google Doc. Developer queries Scopus API and translates result into XML. Use DOI and PMID to lookup record in PubMed.

2. Combine metadata from both sources as a candidate for ingest.

3. If duplicate, disregard. If new, ingest.

4. Re-ingest temporal data such as citation count.
## What is ingested from where?

<table>
<thead>
<tr>
<th>Scopus</th>
<th>PubMed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Author Names</td>
<td>Abstract</td>
</tr>
<tr>
<td>Article Title</td>
<td>Medical Subject Headings (MeSH)</td>
</tr>
<tr>
<td>Journal Title</td>
<td>Funding</td>
</tr>
<tr>
<td>DOI</td>
<td>PubMed Central Identifier</td>
</tr>
<tr>
<td>PMID (PubMed Identifier)</td>
<td>Status (e.g., in process)</td>
</tr>
<tr>
<td>Date of publication</td>
<td>Second ISSN</td>
</tr>
<tr>
<td>ISSN</td>
<td>Language</td>
</tr>
<tr>
<td>Citation count</td>
<td>Journal abbreviation</td>
</tr>
<tr>
<td></td>
<td>Publication type</td>
</tr>
</tbody>
</table>
A key consideration: will a publication ingest be institution-centric or person-centric?
**Query by individual**

### Make Author Selection

<table>
<thead>
<tr>
<th>Author Last Name</th>
<th>Initials or First Name</th>
<th>Show exact matches only</th>
</tr>
</thead>
<tbody>
<tr>
<td>moore</td>
<td>j.p.</td>
<td></td>
</tr>
</tbody>
</table>

**Affiliation**

*E.g., university of toronto*

To determine which author names should be grouped together under a single identifier number, the Scopus Author Identifier uses an algorithm that matches author names based on their affiliation, address, subject area, source title, dates of publication, citations, and co-authors. Documents with insufficient data may not be matched, this can lead to more than one entry in the results list for the same author. By default, only details pages matched to more than one document in Scopus are shown in search results. About Scopus Author Identifier

### Author results: 60 of 174

**Show Profile Matches with One Document**

<table>
<thead>
<tr>
<th>Authors</th>
<th>Documents</th>
<th>Subject Area</th>
<th>Affiliation</th>
<th>City</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moore, John P.</td>
<td>258</td>
<td>Immunology and Microbiology; Medicine; Biochemistry, Genetics and Molecular Biology; ...</td>
<td>Weill Cornell Medical College</td>
<td>New York</td>
</tr>
<tr>
<td>Moore, David J P</td>
<td>176</td>
<td>Medicine; Psychology; Neuroscience; ...</td>
<td>University of Arizona</td>
<td>Tucson</td>
</tr>
<tr>
<td>Duke, Brian O L P J</td>
<td>131</td>
<td>Immunology and Microbiology; Medicine; Mathematics; ...</td>
<td>River Blindness Foundation</td>
<td>Lancaster</td>
</tr>
<tr>
<td>Moore, Philip J.</td>
<td>109</td>
<td>Engineering; Energy; ...</td>
<td>Electric Power Research Institute</td>
<td>Palo Alto</td>
</tr>
</tbody>
</table>

---

*Collective Insight*  *Driven by Shared Data*  *#InsightSeries*  *OCLC*
Query by institution

Make Affiliation Selection

Affiliation

- weill OR (cornell AND medical)
- E.g., university of toronto

The Scopus Affiliation Identifier is the world’s first tool to help you identify and group an organization’s complete body of work. It turns a time-consuming process into a simple task.

About Scopus Affiliation Identifier...

Affiliation results: 8

<table>
<thead>
<tr>
<th>Affiliations</th>
<th>Documents</th>
<th>City</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weill Cornell Medical College</td>
<td>29599</td>
<td>New York</td>
<td>United</td>
</tr>
<tr>
<td>Cornell University Medical College</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cornell University</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Find potential affiliation matches</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weill Cornell Medical Center</td>
<td>8339</td>
<td>New York</td>
<td>United</td>
</tr>
<tr>
<td>Cornell Medical Center</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cornell University Medical Center</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Find potential affiliation matches</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weill Graduate School of Medical Sciences</td>
<td>966</td>
<td>New York</td>
<td>United</td>
</tr>
<tr>
<td>Find potential affiliation matches</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weill Cornell Medical College in Qatar</td>
<td>190</td>
<td>Doha</td>
<td>Qatar</td>
</tr>
<tr>
<td>Find potential affiliation matches</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weill Cornell Medical College in Qatar</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Find potential affiliation matches</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Cornell Institute of Geriatric Psychiatry</td>
<td>60</td>
<td>New York</td>
<td>United</td>
</tr>
<tr>
<td>Cornell University</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cornell Inst. of Geriatric Psychiat.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Find potential affiliation matches</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weill Bugando University College of Health Sciences</td>
<td>40</td>
<td>Mwanza</td>
<td>Tanzania</td>
</tr>
<tr>
<td>Find potential affiliation matches</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weill Bugando University College of Health Sciences</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Refine results

- Limit to
- Exclude

City

- New York (6)
- Doha (1)
- Mwanza (1)

Country

- United States (6)
- Qatar (1)
- Tanzania (1)
Query by institution
Affiliation ID = “Weill”

- Easier to identify hits
- Easier for institutional reporting, especially year to year comparisons
- Assertions of co-author identity can be unclear

Query by person
Author ID = “8256757” x 1300

- More laborious – need an internal source for people
- Often accounts for publications w/ no or incorrect affiliation
- Accounts for previous affiliations
Scopus commits two varieties of disambiguation errors

**Splitting** - one person, multiple author IDs; relatively easy to recover from

**Lumping** - multiple people, one author ID; relatively hard to recover from
<table>
<thead>
<tr>
<th>#</th>
<th>Last Name</th>
<th>First Name</th>
<th>Affiliation</th>
<th>Institution</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>Brown</td>
<td>A. M C</td>
<td>Show Last Title</td>
<td>Show Last Title</td>
<td>Show Last Title</td>
</tr>
<tr>
<td>13</td>
<td>Brown</td>
<td>A. M C</td>
<td>Show Last Title</td>
<td>Show Last Title</td>
<td>Show Last Title</td>
</tr>
<tr>
<td>14</td>
<td>Brown</td>
<td>A. M C</td>
<td>Show Last Title</td>
<td>Show Last Title</td>
<td>Show Last Title</td>
</tr>
<tr>
<td>15</td>
<td>Brown</td>
<td>A. M C</td>
<td>Show Last Title</td>
<td>Show Last Title</td>
<td>Show Last Title</td>
</tr>
<tr>
<td>16</td>
<td>Brown</td>
<td>A. M C</td>
<td>Show Last Title</td>
<td>Show Last Title</td>
<td>Show Last Title</td>
</tr>
<tr>
<td>17</td>
<td>Brown</td>
<td>A. M C</td>
<td>Show Last Title</td>
<td>Show Last Title</td>
<td>Show Last Title</td>
</tr>
<tr>
<td>18</td>
<td>Brown</td>
<td>A. M C</td>
<td>Show Last Title</td>
<td>Show Last Title</td>
<td>Show Last Title</td>
</tr>
<tr>
<td>19</td>
<td>Brown</td>
<td>A. M C</td>
<td>Show Last Title</td>
<td>Show Last Title</td>
<td>Show Last Title</td>
</tr>
<tr>
<td>20</td>
<td>Brown</td>
<td>A. M C</td>
<td>Show Last Title</td>
<td>Show Last Title</td>
<td>Show Last Title</td>
</tr>
<tr>
<td>21</td>
<td>Brown</td>
<td>A. M C</td>
<td>Show Last Title</td>
<td>Show Last Title</td>
<td>Show Last Title</td>
</tr>
</tbody>
</table>

**Example of splitting**
### Example of lumping

Either this person is very cavalier about the spelling of their name, or there's more than one M. Cohen.
How accurate is Scopus at author disambiguation c. 2013?

Gold standard = librarian judgment

- **Ideal** – one-to-one relation between Scopus author ID and person
  - n=369

- **Splitting** – more than one author ID per person
  - n=707

- **Lumping** – more than one person per author ID
  - n=86

- **Both errors**
  - n=23
Two author disambiguation methods against a gold standard

From Johnson et al. Submitted. “Automatic generation of investigator bibliographies for institutional research networking systems.”
“Special queries” can compensate for lumping errors

Query #1
Cornell University OR Weill Cornell Medical College
OR Weill Cornell Medical Center OR New York
University School of Medicine OR Aaron Diamond
AIDS Research Center OR Institute of Cancer
Research London OR University of Glasgow OR
University of Cambridge

Query #2
Moore, John P.

<table>
<thead>
<tr>
<th>Query</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>352,948</td>
</tr>
<tr>
<td>#2</td>
<td>399</td>
</tr>
<tr>
<td>#2 ∩ #1</td>
<td>239</td>
</tr>
</tbody>
</table>

Actually three people: a South African plant biologist, an Australian IT expert, and our WCMC-based immunologist.
Examples of special queries

(AU-ID(7405920800))
AND
(AF-ID(60007997) OR AF-ID(60009470) OR AF-ID(60019868))

(AU-ID(7402763146))
AND
(AF-ID(60007997) OR AF-ID(60019868) OR AF-ID(60018043) OR
AF-ID(60007997) OR AF-ID(60019868) OR AF-ID(100366692) OR
AF-ID(60018043) OR AF-ID(60002339) OR AF-ID(60009343) OR
AF-ID(60024541) OR AF-ID(60025843) OR AF-ID(60027565))
How can VIVO data address pressing institutional needs in order to strengthen its viability?
NIH Open Access policy compliance

WCMC authors who have received NIH funding but haven’t deposited pre-prints in PubMed Central receive a nastygram personalized notice.
Co-author network and expertise of arbitrary group of faculty
Suggested publications in annual faculty review tool
Administrators are avid consumers of institutional data.
Administrators want reporting tools (especially about publications) that are:

- Have current data
- Easy to use
- Allow for sophisticated queries
VIVO Dashboard now under development

VIVO Dashboard
A Drupal-based tool for visualizing semantic data

Publications
The following publications are for all publications by active Weill Cornell Medical College faculty as represented in VIVO.
Expertise recommendation tool also under development
Acknowledgements

Eliza Chan and Prakash Adekkanattu - developers at Weill Cornell

Don Carpenter and Zeheng Wang - VIVO Dashboard developer

Jie Lin - Expertise Recommendation Tool developer

Drew Wright - publications help and NIH Access Policy compliance