# Integrating and exploiting public metadata sources in a bibliographic information system

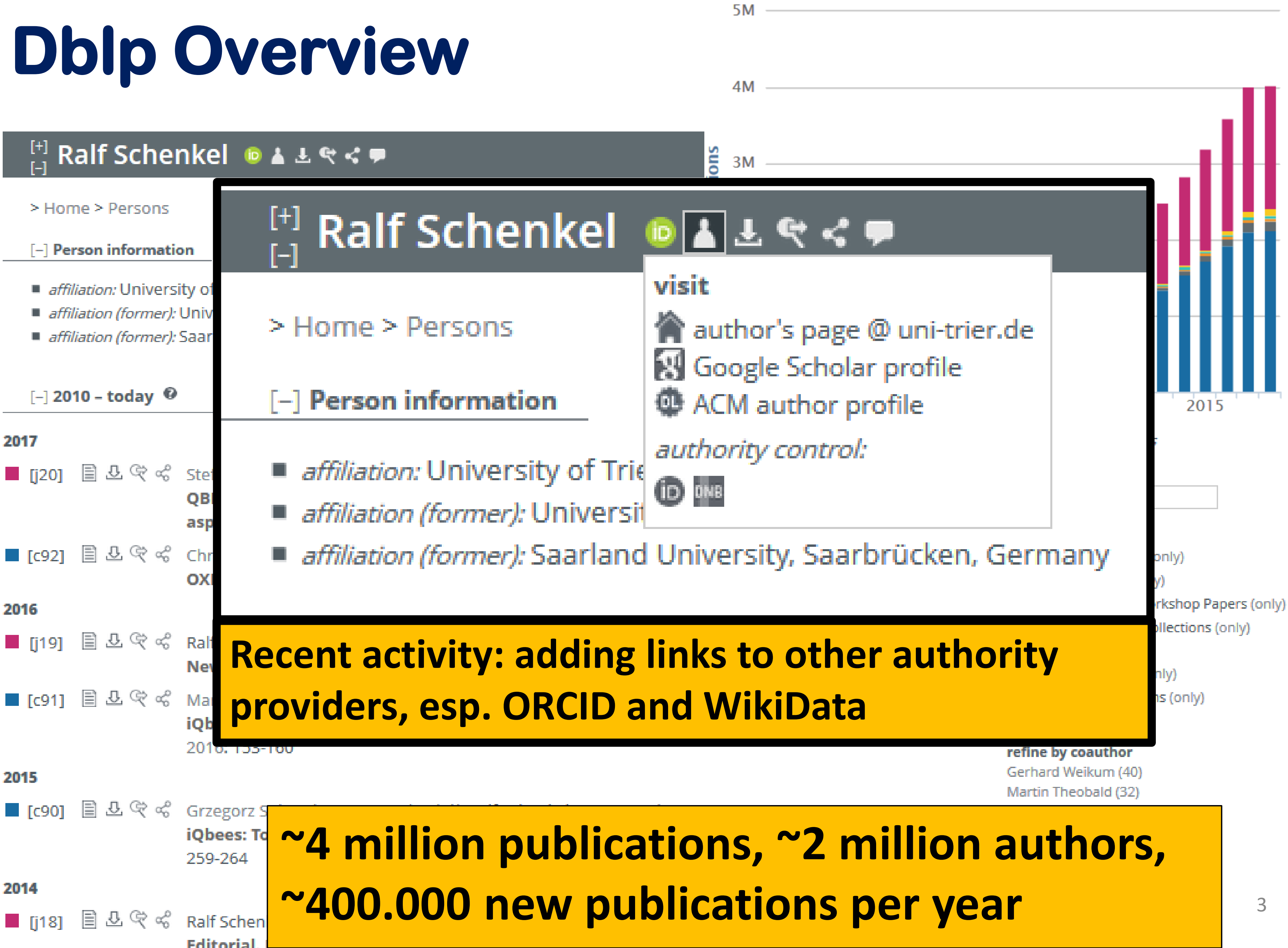
*Ralf Schenkel*

dblp
computer science bibliography

CIRT

Universität Trier

# Dblp Overview



[+]
[–] Ralf Schenkel

> Home > Persons

[–] Person information

- *affiliation:* University of Trie
- *affiliation (former):* Universit
- *affiliation (former):* Saarland University, Saarbrücken, Germany

visit

- author's page @ uni-trier.de
- Google Scholar profile
- ACM author profile

*authority control:*

[–] 2010 – today

2017

[j20] Ste QB asp

[c92] Chr OX

2016

[j19] Ral Ne

[c91] Ma iQb 2016: 153-160

2015

[c90] Grzegorz S iQbees: To 259-264

2014

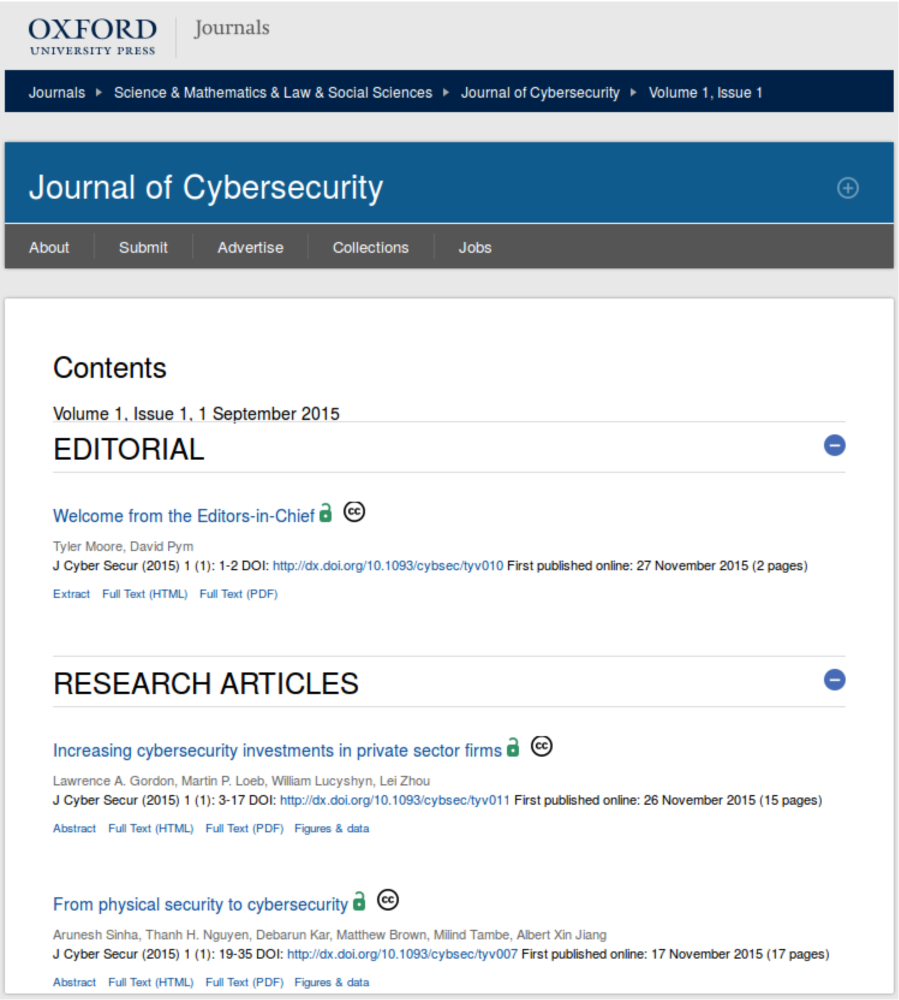
[j18] Ralf Schen Editorial

refine by coauthor
Gerhard Weikum (40)
Martin Theobald (32)

**Recent activity: adding links to other authority providers, esp. ORCID and WikiData**

**~4 million publications, ~2 million authors, ~400.000 new publications per year**

# How is publication data added to dblp?



OXFORD UNIVERSITY PRESS | Journals

Journals ▸ Science & Mathematics & Law & Social Sciences ▸ Journal of Cybersecurity ▸ Volume 1, Issue 1

Journal of Cybersecurity

About | Submit | Advertise | Collections | Jobs

Contents

Volume 1, Issue 1, 1 September 2015

EDITORIAL

Welcome from the Editors-in-Chief

Tyler Moore, David Pym
J Cyber Secur (2015) 1 (1): 1-2 DOI: http://dx.doi.org/10.1093/cybsec/tyv010 First published online: 27 November 2015 (2 pages)

Extract | Full Text (HTML) | Full Text (PDF)

RESEARCH ARTICLES

Increasing cybersecurity investments in private sector firms

Lawrence A. Gordon, Martin P. Loeb, William Lucyshyn, Lei Zhou
J Cyber Secur (2015) 1 (1): 3-17 DOI: http://dx.doi.org/10.1093/cybsec/tyv011 First published online: 26 November 2015 (15 pages)

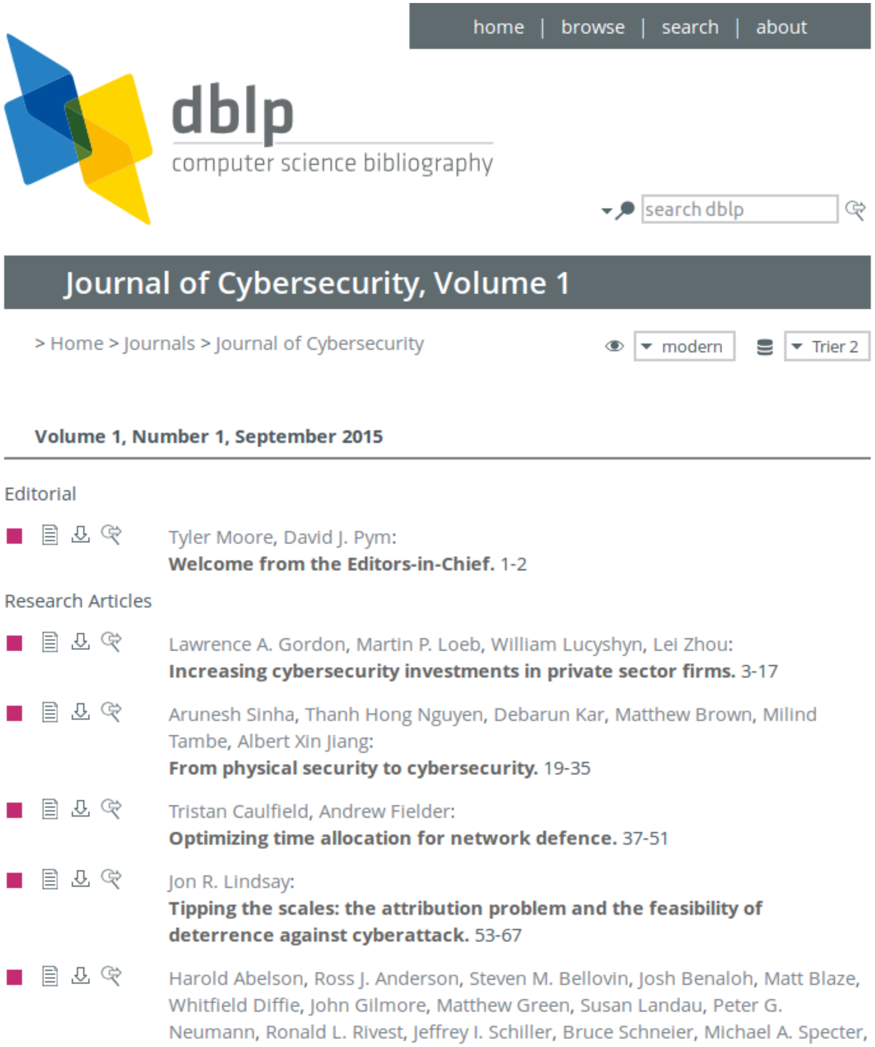
Abstract | Full Text (HTML) | Full Text (PDF) | Figures & data

From physical security to cybersecurity

Arunesh Sinha, Thanh H. Nguyen, Debarun Kar, Matthew Brown, Milind Tambe, Albert Xin Jiang
J Cyber Secur (2015) 1 (1): 19-35 DOI: http://dx.doi.org/10.1093/cybsec/tyv007 First published online: 17 November 2015 (17 pages)

Abstract | Full Text (HTML) | Full Text (PDF) | Figures & data

?



home | browse | search | about

dblp
computer science bibliography

search dblp

Journal of Cybersecurity, Volume 1

> Home > Journals > Journal of Cybersecurity

modern | Trier 2

Volume 1, Number 1, September 2015

Editorial

Tyler Moore, David J. Pym:
**Welcome from the Editors-in-Chief.** 1-2

Research Articles

Lawrence A. Gordon, Martin P. Loeb, William Lucyshyn, Lei Zhou:
**Increasing cybersecurity investments in private sector firms.** 3-17

Arunesh Sinha, Thanh Hong Nguyen, Debarun Kar, Matthew Brown, Milind Tambe, Albert Xin Jiang:
**From physical security to cybersecurity.** 19-35

Tristan Caulfield, Andrew Fielder:
**Optimizing time allocation for network defence.** 37-51

Jon R. Lindsay:
**Tipping the scales: the attribution problem and the feasibility of deterrence against cyberattack.** 53-67

Harold Abelson, Ross J. Anderson, Steven M. Bellovin, Josh Benaloh, Matt Blaze, Whitfield Diffie, John Gilmore, Matthew Green, Susan Landau, Peter G. Neumann, Ronald L. Rivest, Jeffrey I. Schiller, Bruce Schneier, Michael A. Specter,

# Dblp Data Ingestion Pipeline

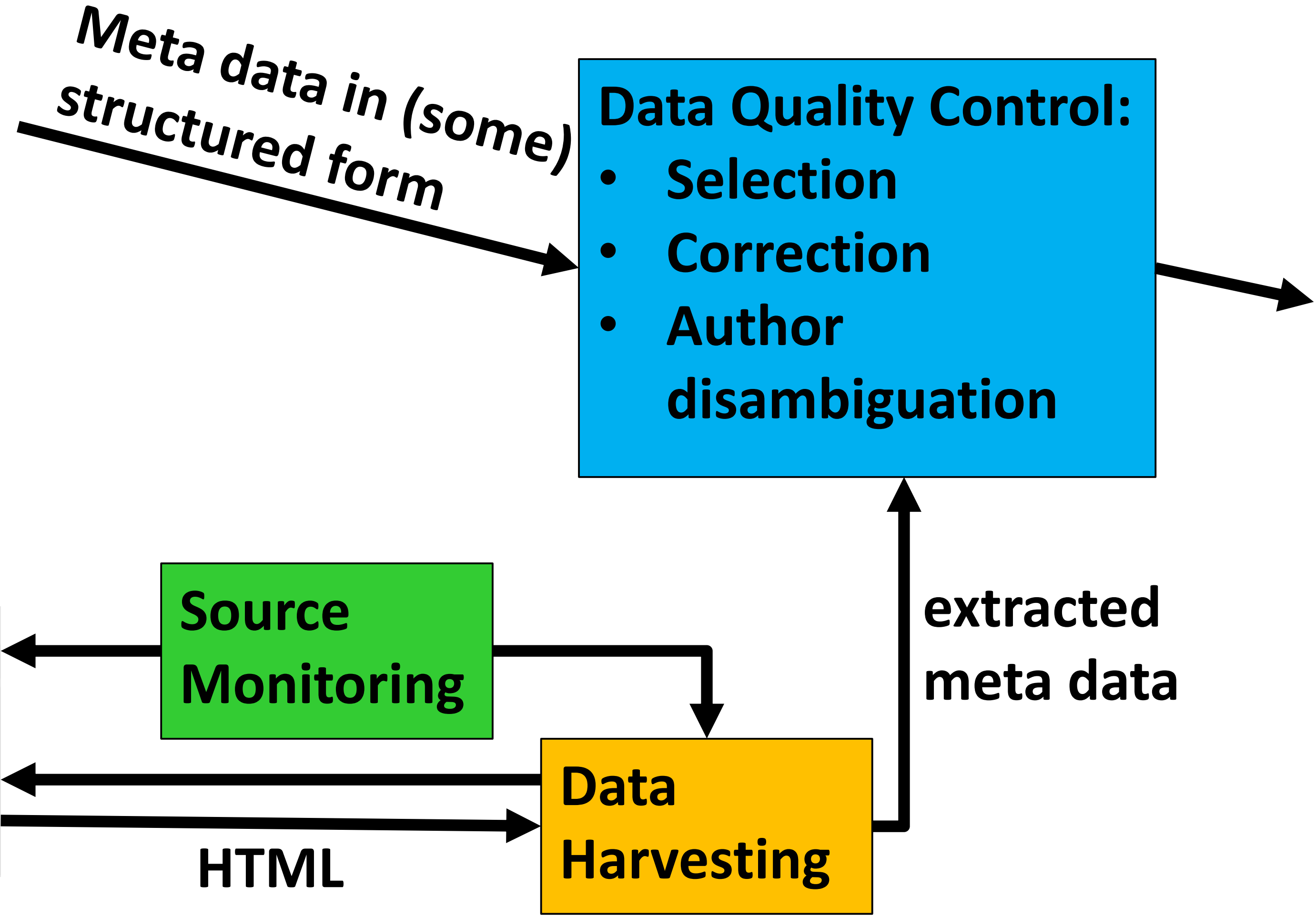
publishers

Springer

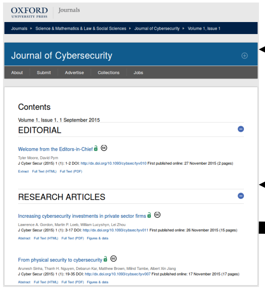
IEEE

ELSEVIER

acm



Web



OXFORD UNIVERSITY PRESS | Journals

Journals ▸ Science & Mathematics & Law & Social Sciences ▸ Journal of Cybersecurity ▸ Volume 1, Issue 1

Journal of Cybersecurity

Contents

Volume 1, Issue 1, 1 September 2015

EDITORIAL

Welcome from the Editors-in-Chief

Tyler Moore, David Pym

J Cyber Secur (2015) 1 (1): 1-2 DOI: http://dx.doi.org/10.1093/cybsec/tyv010 First published online: 27 November 2015 (2 pages)

Extract Full Text (HTML) Full Text (PDF)

RESEARCH ARTICLES

Increasing cybersecurity investments in private sector firms

Lawrence A. Gordon, Martin P. Loeb, William Lucyshyn, Lei Zhou

J Cyber Secur (2015) 1 (1): 3-17 DOI: http://dx.doi.org/10.1093/cybsec/tyv011 First published online: 26 November 2015 (15 pages)

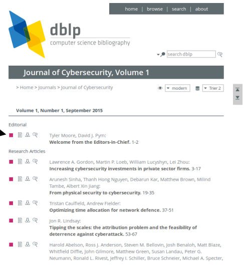
Abstract Full Text (HTML) Full Text (PDF) Figures & data

From physical security to cybersecurity

Arunesh Sinha, Thanh H. Nguyen, Debarun Kar, Matthew Brown, Milind Tambe, Albert Xin Jiang

J Cyber Secur (2015) 1 (1): 19-35 DOI: http://dx.doi.org/10.1093/cybsec/tyv007 First published online: 17 November 2015 (17 pages)

Abstract Full Text (HTML) Full Text (PDF) Figures & data



Journal of Cybersecurity, Volume 1

> Home > Journals > Journal of Cybersecurity

Volume 1, Number 1, September 2015

Editorial

Tyler Moore, David J. Pym:
**Welcome from the Editors-in-Chief.** 1-2

Research Articles

Lawrence A. Gordon, Martin P. Loeb, William Lucyshyn, Lei Zhou:
**Increasing cybersecurity investments in private sector firms.** 3-17

Arunesh Sinha, Thanh Hong Nguyen, Debarun Kar, Matthew Brown, Milind Tambe, Albert Xin Jiang:
**From physical security to cybersecurity.** 19-35

Tristan Caulfield, Andrew Fielder:
**Optimizing time allocation for network defence.** 37-51

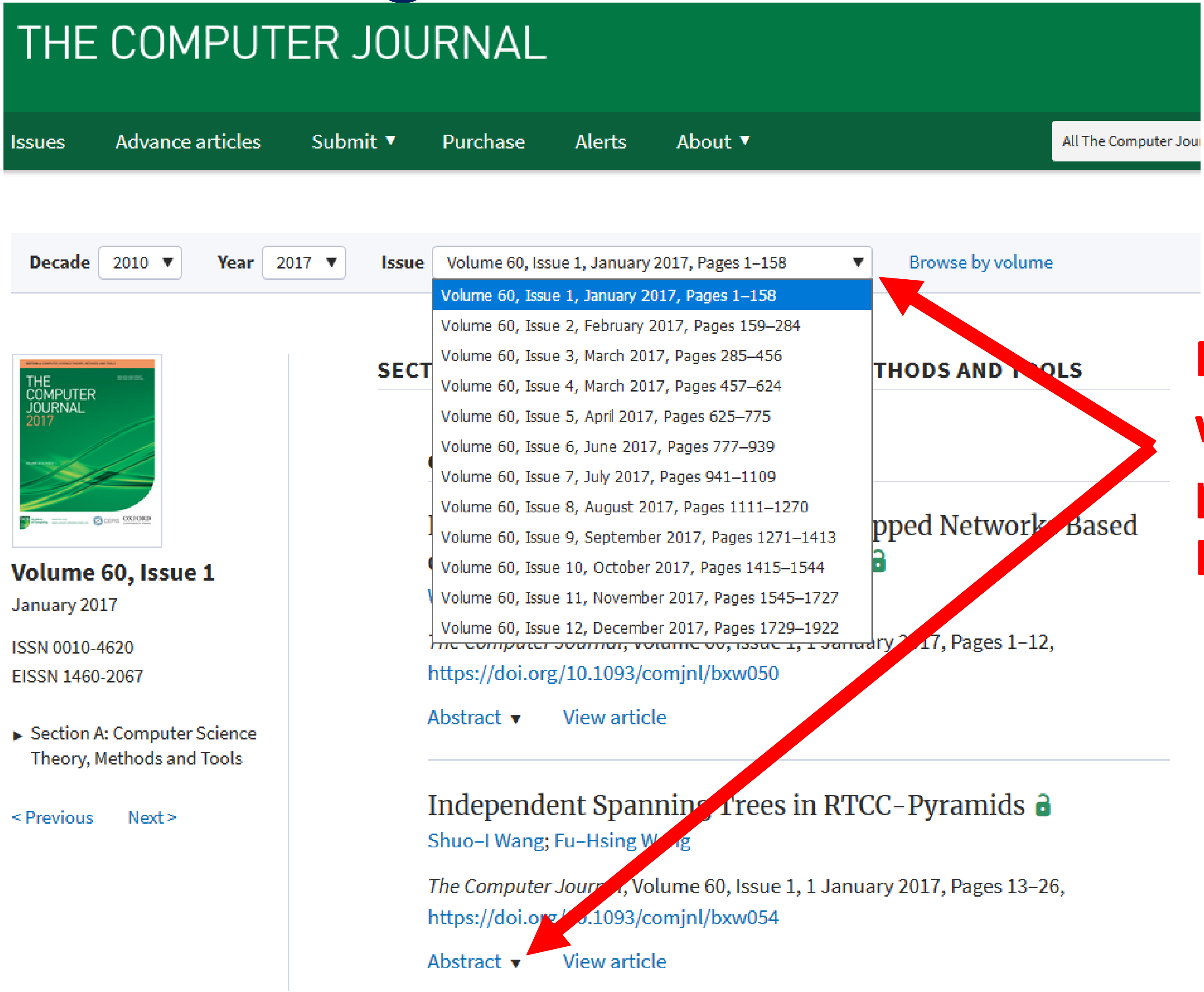
Jon R. Lindsay:
**Tipping the scales: the attribution problem and the feasibility of deterrence against cyberattack.** 53-67

Harold Abelson, Ross J. Anderson, Steven M. Bellovin, Josh Benaloh, Matt Blaze, Whitfield Diffie, John Gilmore, Matthew Green, Susan Landau, Peter G. Neumann, Ronald L. Rivest, Jeffrey I. Schiller, Bruce Schneier, Michael A. Specter,

# Outline

- **Meta Data Harvesting**
- Author Disambiguation
- Existing Metadata Collections
- Citations

# Harvesting is much more difficult now



THE COMPUTER JOURNAL

Issues | Advance articles | Submit ▼ | Purchase | Alerts | About ▼ | All The Computer Jou

**Decade** 2010 ▼ **Year** 2017 ▼ **Issue** Volume 60, Issue 1, January 2017, Pages 1–158 ▼ Browse by volume

- Volume 60, Issue 1, January 2017, Pages 1–158
- Volume 60, Issue 2, February 2017, Pages 159–284
- Volume 60, Issue 3, March 2017, Pages 285–456
- Volume 60, Issue 4, March 2017, Pages 457–624
- Volume 60, Issue 5, April 2017, Pages 625–775
- Volume 60, Issue 6, June 2017, Pages 777–939
- Volume 60, Issue 7, July 2017, Pages 941–1109
- Volume 60, Issue 8, August 2017, Pages 1111–1270
- Volume 60, Issue 9, September 2017, Pages 1271–1413
- Volume 60, Issue 10, October 2017, Pages 1415–1544
- Volume 60, Issue 11, November 2017, Pages 1545–1727
- Volume 60, Issue 12, December 2017, Pages 1729–1922

**Volume 60, Issue 1**

January 2017

ISSN 0010-4620

EISSN 1460-2067

▶ Section A: Computer Science Theory, Methods and Tools

< Previous Next >

SECT … THODS AND … OLS

… pped Network Based

… ary 2017, Pages 1–12,

https://doi.org/10.1093/comjnl/bxw050

Abstract ▼ View article

Independent Spanning Trees in RTCC-Pyramids

Shuo-I Wang; Fu-Hsing Wang

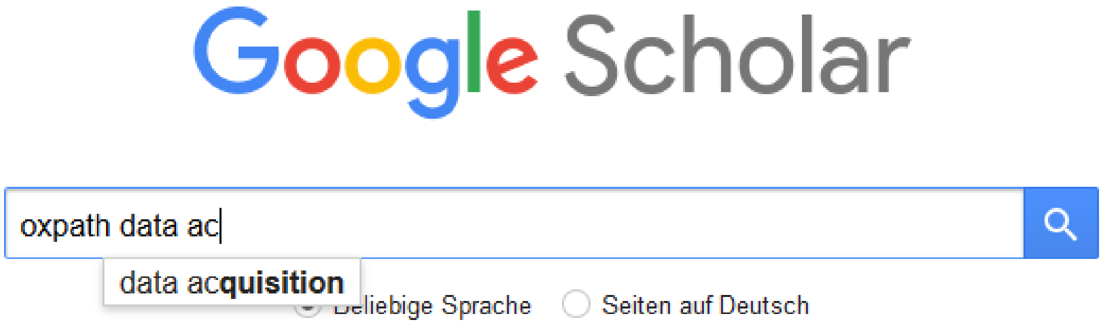
*The Computer Journal*, Volume 60, Issue 1, 1 January 2017, Pages 13–26,

https://doi.org/10.1093/comjnl/bxw054

Abstract ▼ View article

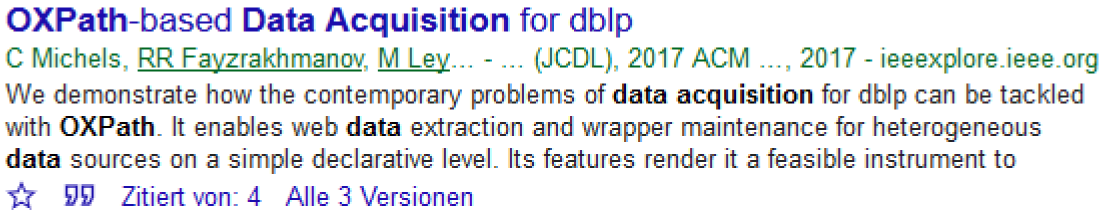
**Need to interact with Web site, parsing static HTML not enough**

# Harvesting is much more difficult now



oxpath data ac

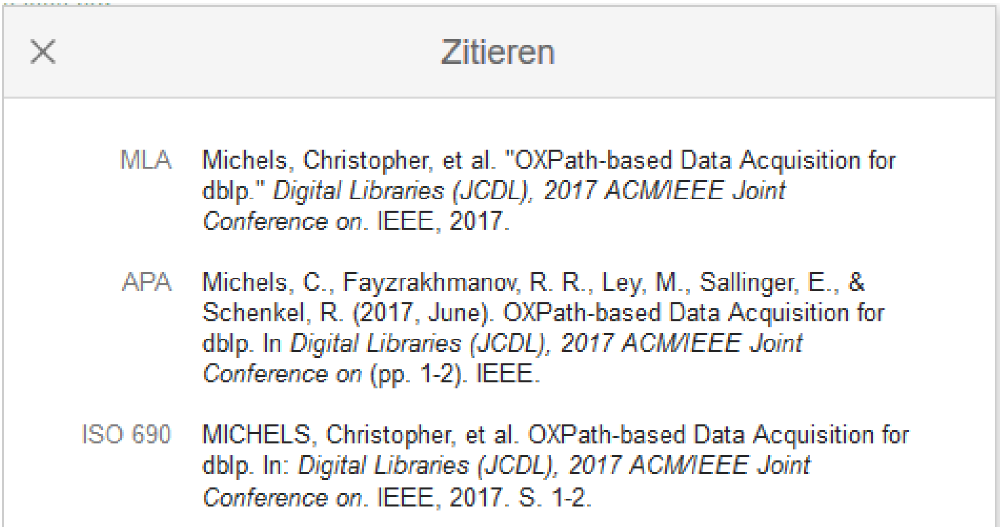
data acquisition



**OXPath**-based **Data Acquisition** for dblp

C Michels, RR Fayzrakhmanov, M Ley… - … (JCDL), 2017 ACM …, 2017 - ieeexplore.ieee.org

We demonstrate how the contemporary problems of **data acquisition** for dblp can be tackled with **OXPath**. It enables web **data** extraction and wrapper maintenance for heterogeneous **data** sources on a simple declarative level. Its features render it a feasible instrument to

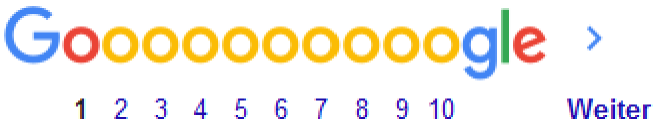


Zitieren

MLA Michels, Christopher, et al. "OXPath-based Data Acquisition for dblp." *Digital Libraries (JCDL), 2017 ACM/IEEE Joint Conference on*. IEEE, 2017.

APA Michels, C., Fayzrakhmanov, R. R., Ley, M., Sallinger, E., & Schenkel, R. (2017, June). OXPath-based Data Acquisition for dblp. In *Digital Libraries (JCDL), 2017 ACM/IEEE Joint Conference on* (pp. 1-2). IEEE.

ISO 690 MICHELS, Christopher, et al. OXPath-based Data Acquisition for dblp. In: *Digital Libraries (JCDL), 2017 ACM/IEEE Joint Conference on*. IEEE, 2017. S. 1-2.



**Successful harvesting needs to implement Javascript**

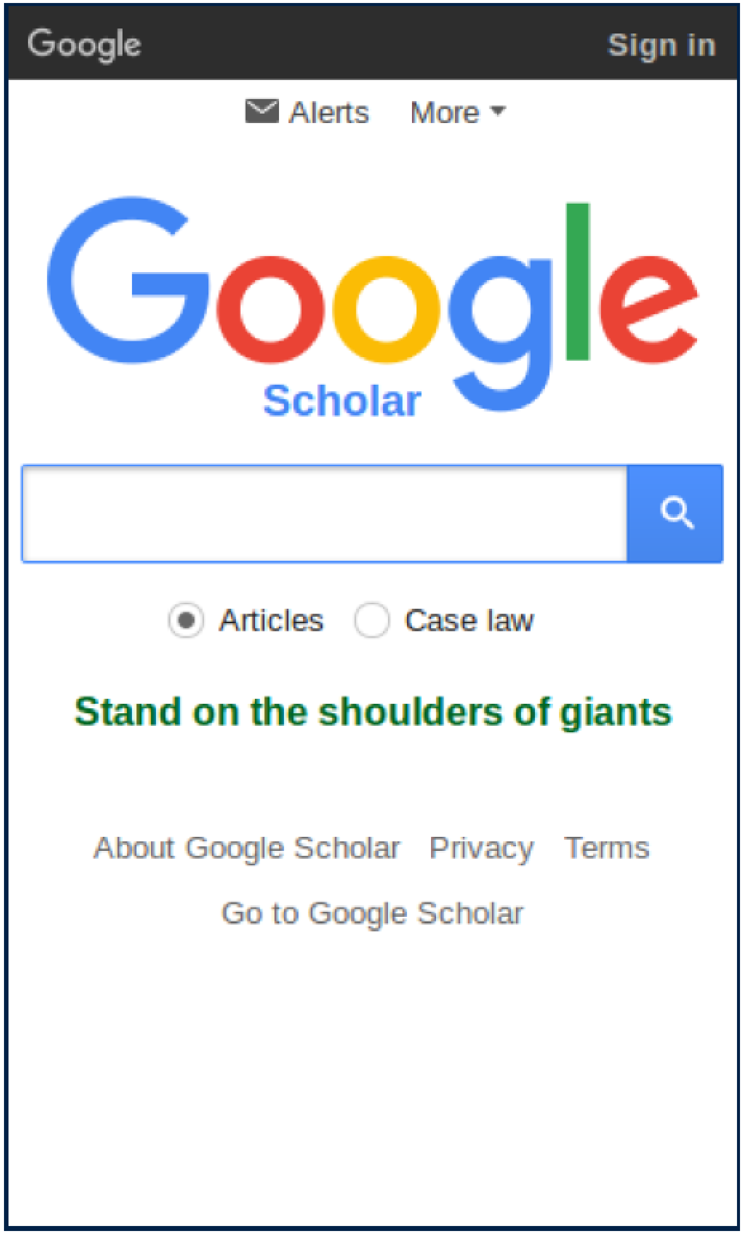
# Monitoring & harvesting: OXPath

Extension of XPath by University of Oxford (Georg Gottlob et al.)

- Actions: fill in forms, click buttons
- Extraction: specify what should be harvested
- Transformation: specify target XML format
- Iteration: loops, e.g., for paginated content

Michels, C., Fayzrakhmanov, R.R., Ley, M., Sallinger, E., Schenkel, R.: OXPath-based data acquisition for dblp. In: 2017 ACM/IEEE Joint Conference on Digital Libraries, 2017

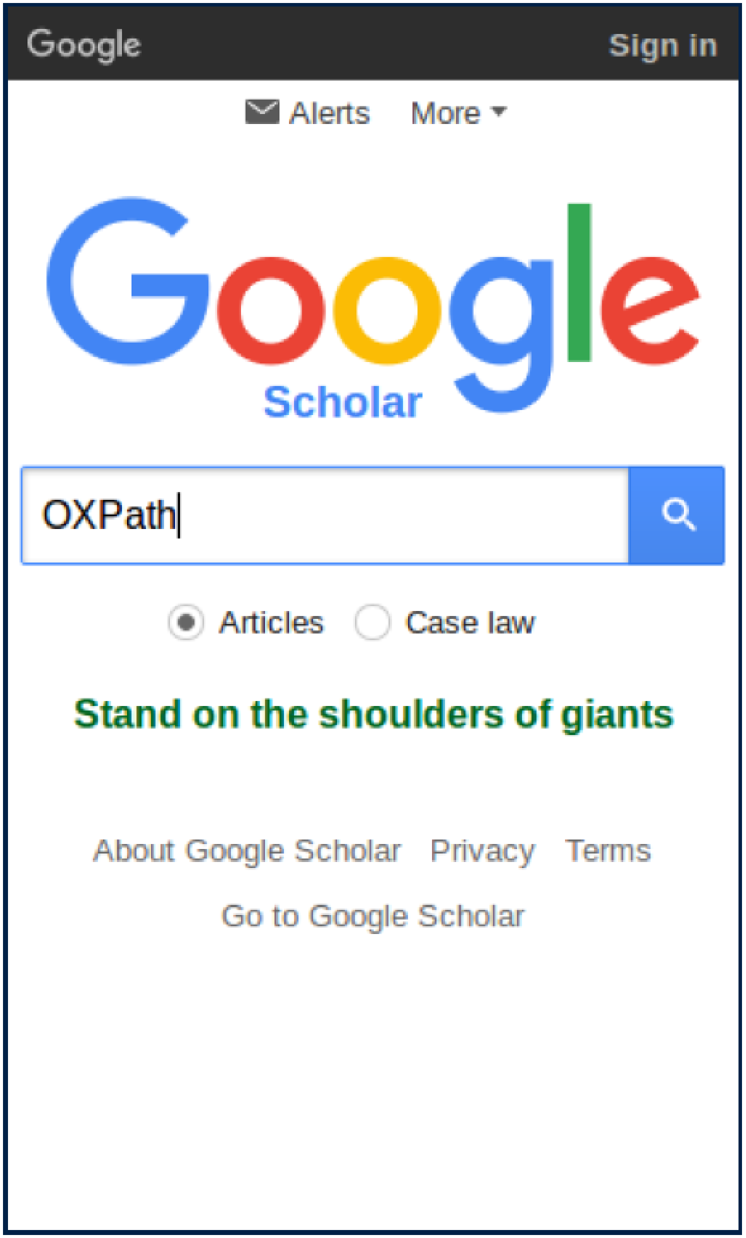
# Example: Navigating Google Scholar



OXPath Expression

```
doc("https://scholar.google.com")
```

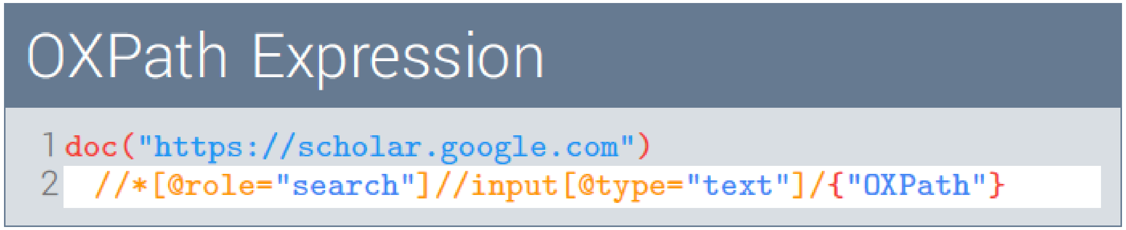
# Example: Navigating Google Scholar



OXPath

● Articles ○ Case law

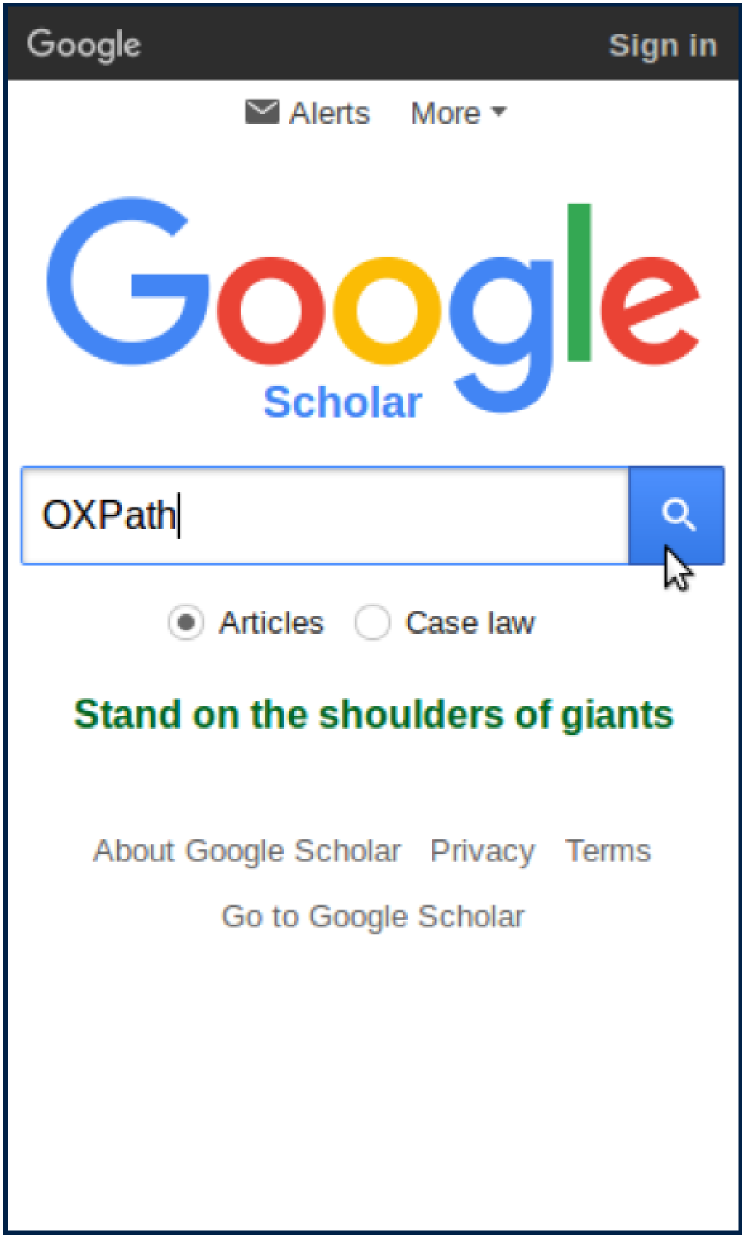
Stand on the shoulders of giants



## OXPath Expression

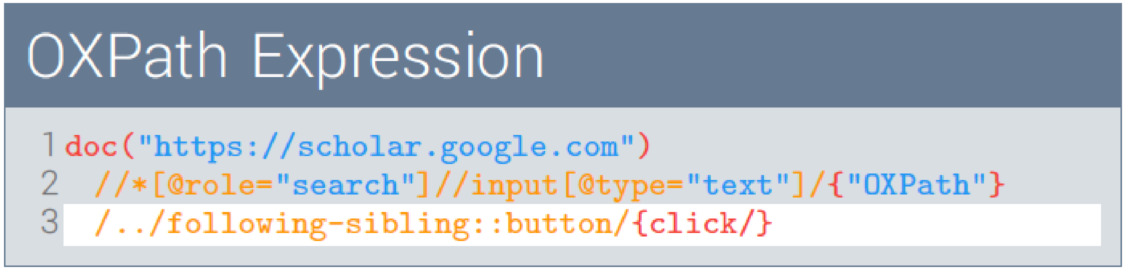
```
doc("https://scholar.google.com")
 //*[@role="search"]//input[@type="text"]/{"OXPath"}
```

# Example: Navigating Google Scholar



OXPath

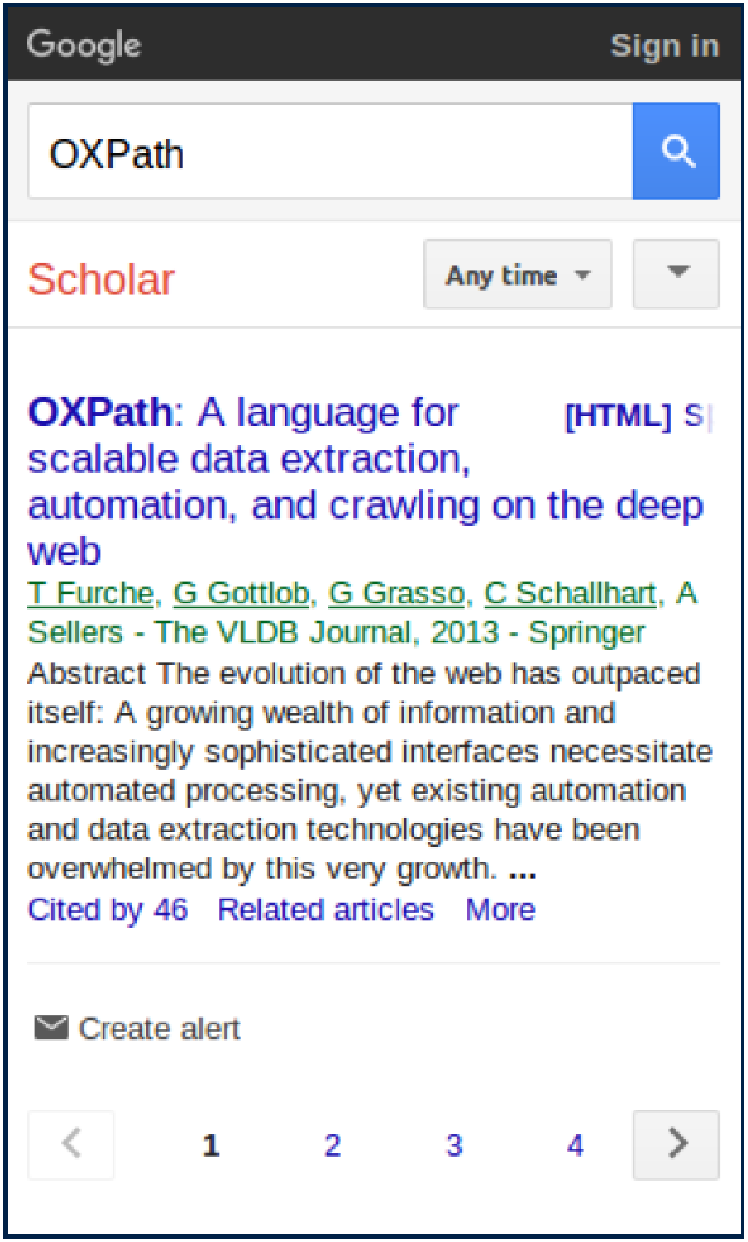
● Articles ○ Case law



OXPath Expression

```
doc("https://scholar.google.com")
 //*[@role="search"]//input[@type="text"]/{"OXPath"}
 /../following-sibling::button/{click/}
```

# Example: Navigating Google Scholar



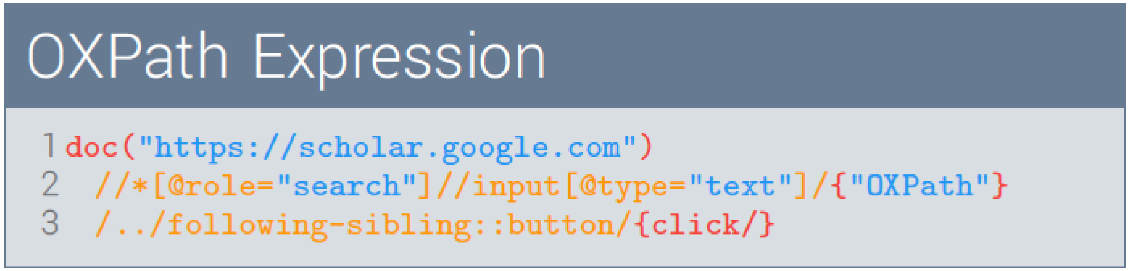
**OXPath**: A language for scalable data extraction, automation, and crawling on the deep web

T Furche, G Gottlob, G Grasso, C Schallhart, A Sellers - The VLDB Journal, 2013 - Springer

Abstract The evolution of the web has outpaced itself: A growing wealth of information and increasingly sophisticated interfaces necessitate automated processing, yet existing automation and data extraction technologies have been overwhelmed by this very growth. **...**

Cited by 46 Related articles More

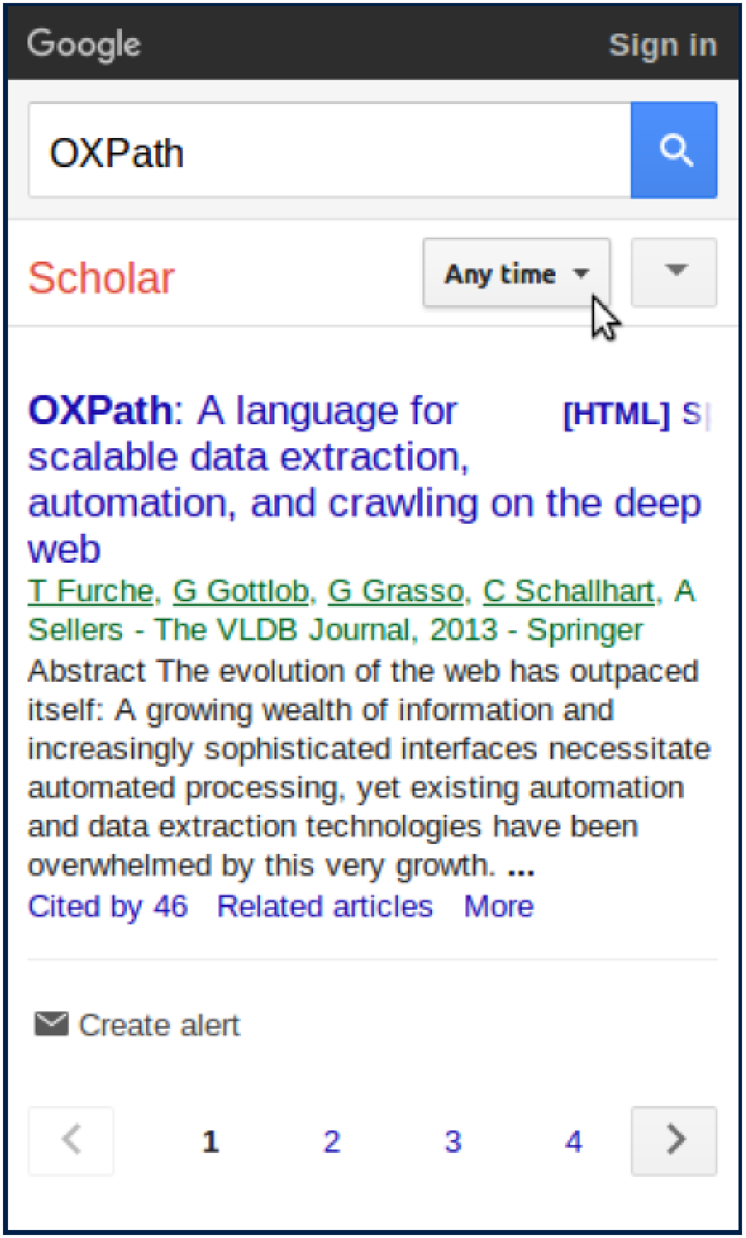
Create alert
1
2
3
4



## OXPath Expression

```
doc("https://scholar.google.com")
 //*[@role="search"]//input[@type="text"]/{"OXPath"}
 /../following-sibling::button/{click/}
```

# Example: Navigating Google Scholar



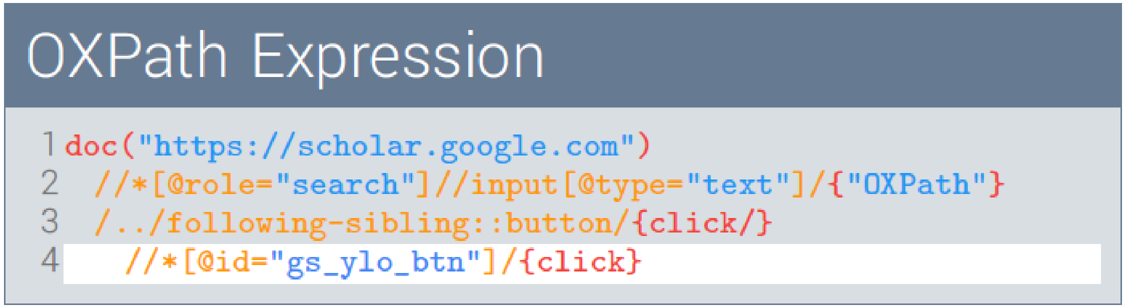
OXPath

**OXPath**: A language for scalable data extraction, automation, and crawling on the deep web **[HTML] S**

T Furche, G Gottlob, G Grasso, C Schallhart, A Sellers - The VLDB Journal, 2013 - Springer

Abstract The evolution of the web has outpaced itself: A growing wealth of information and increasingly sophisticated interfaces necessitate automated processing, yet existing automation and data extraction technologies have been overwhelmed by this very growth. **...**

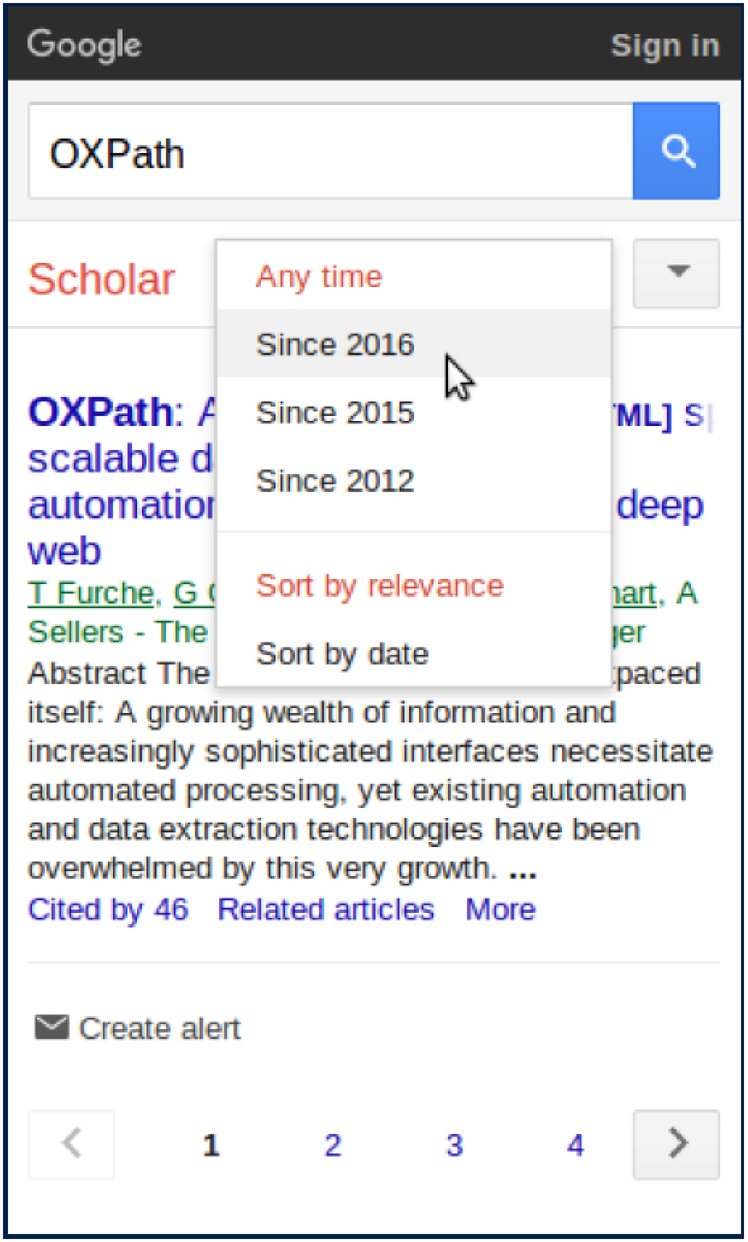
Cited by 46 Related articles More



OXPath Expression

```
doc("https://scholar.google.com")
  //*[@role="search"]//input[@type="text"]/{"OXPath"}
  /../following-sibling::button/{click/}
    //*[@id="gs_ylo_btn"]/{click}
```

# Example: Navigating Google Scholar

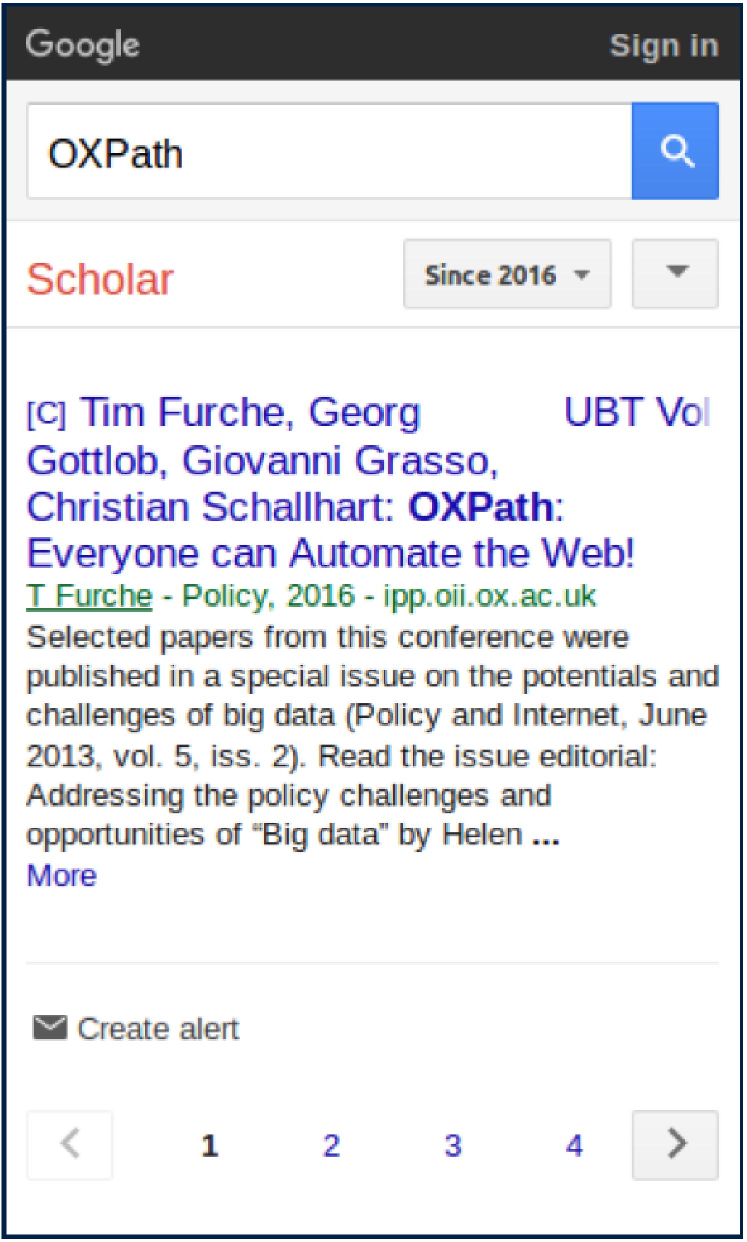


OXPath: A ... ML] S
scalable d ... automatior ... deep web
T Furche, G ( ... nart, A Sellers - The ... ger
Abstract The ... paced itself: A growing wealth of information and increasingly sophisticated interfaces necessitate automated processing, yet existing automation and data extraction technologies have been overwhelmed by this very growth. ...
Cited by 46 Related articles More

## OXPath Expression

```
doc("https://scholar.google.com")
 //*[@role="search"]//input[@type="text"]/{"OXPath"}
 /../following-sibling::button/{click/}
   //*[@id="gs_ylo_btn"]/{click}
     //following::*[@id="gs_ylo_md"]/a[contains(.,
         "2016")]/{click/}
```

# Example: Navigating Google Scholar



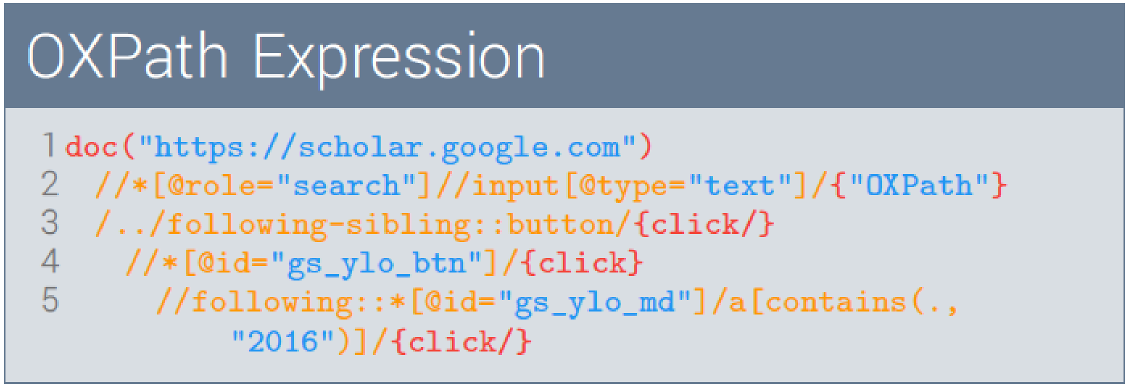
[C] Tim Furche, Georg Gottlob, Giovanni Grasso, Christian Schallhart: **OXPath**: Everyone can Automate the Web! UBT Vol

T Furche - Policy, 2016 - ipp.oii.ox.ac.uk

Selected papers from this conference were published in a special issue on the potentials and challenges of big data (Policy and Internet, June 2013, vol. 5, iss. 2). Read the issue editorial: Addressing the policy challenges and opportunities of "Big data" by Helen **...**

More

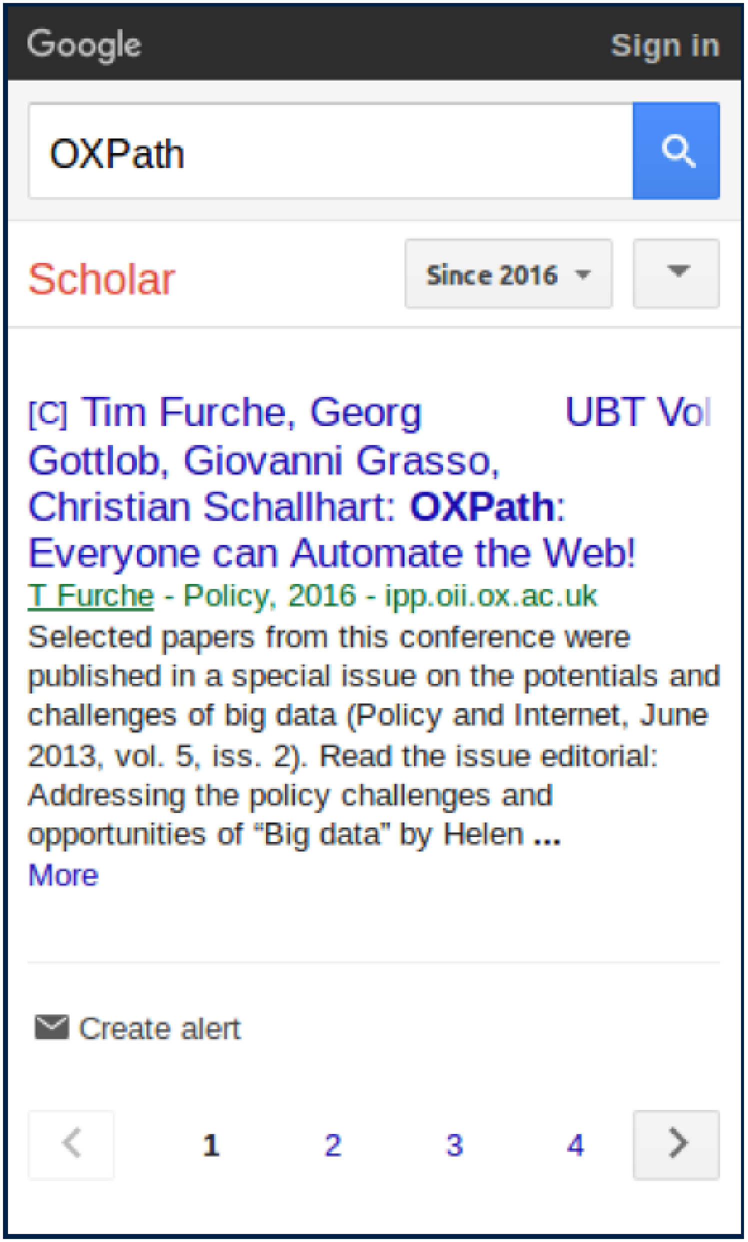
Create alert
‹
1
2
3
4
›



## OXPath Expression

```
doc("https://scholar.google.com")
 //*[@role="search"]//input[@type="text"]/{"OXPath"}
 /../following-sibling::button/{click/}
   //*[@id="gs_ylo_btn"]/{click}
     //following::*[@id="gs_ylo_md"]/a[contains(.,
          "2016")]/{click/}
```

# Example: Navigating Google Scholar



Google Sign in

OXPath

Scholar Since 2016

[C] Tim Furche, Georg Gottlob, Giovanni Grasso, Christian Schallhart: **OXPath**: Everyone can Automate the Web! UBT Vol

T Furche - Policy, 2016 - ipp.oii.ox.ac.uk

Selected papers from this conference were published in a special issue on the potentials and challenges of big data (Policy and Internet, June 2013, vol. 5, iss. 2). Read the issue editorial: Addressing the policy challenges and opportunities of "Big data" by Helen **...**

More

Create alert

1 2 3 4

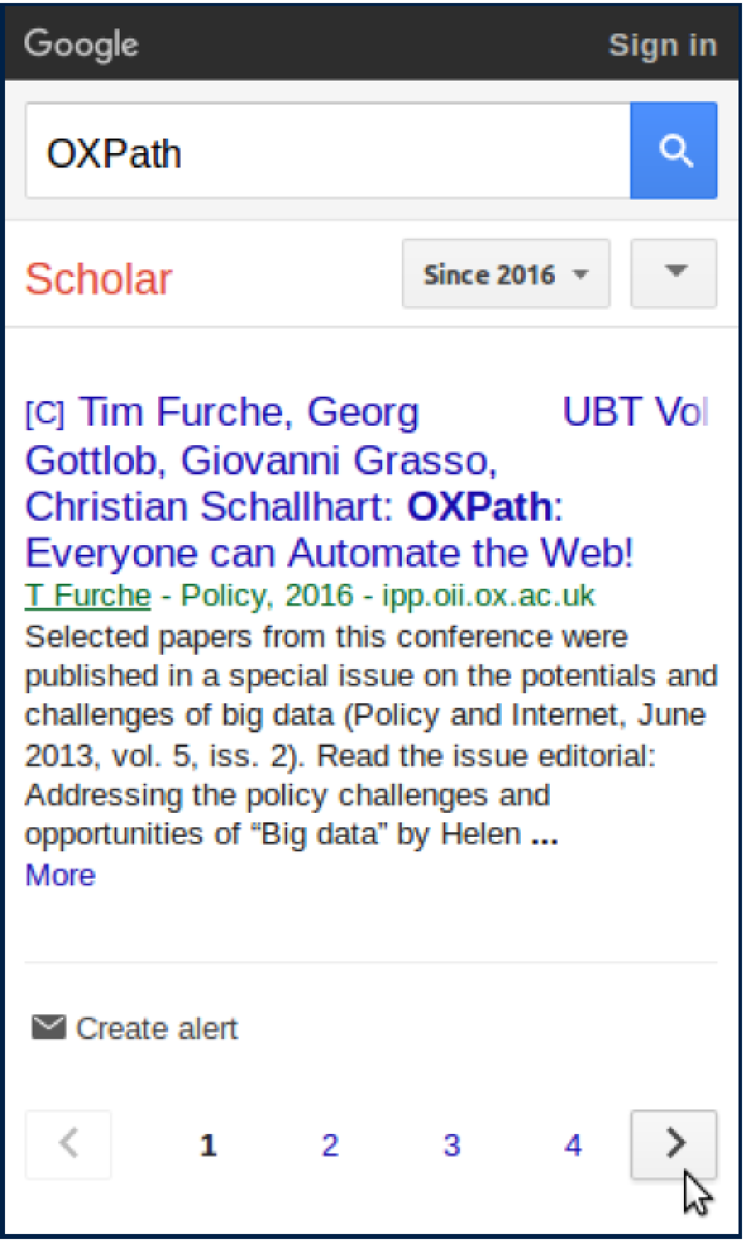
## OXPath Expression

```
doc("https://scholar.google.com")
 //*[@role="search"]//input[@type="text"]/{"OXPath"}
 /../following-sibling::button/{click/}
   //*[@id="gs_ylo_btn"]/{click}
     //following::*[@id="gs_ylo_md"]/a[contains(.,
         "2016")]/{click/}
   //div[@class="gs_ri"]//h3/a:<title=string(.)>
```

## XML Output

```
<?xml version="1.0" encoding="UTF-8"?>
<results>
  <title>Tim Furche, Georg Gottlob, [...]</title>
</results>
```

# Example: Navigating Google Scholar



[C] Tim Furche, Georg Gottlob, Giovanni Grasso, Christian Schallhart: **OXPath**: Everyone can Automate the Web! UBT Vol

T Furche - Policy, 2016 - ipp.oii.ox.ac.uk

Selected papers from this conference were published in a special issue on the potentials and challenges of big data (Policy and Internet, June 2013, vol. 5, iss. 2). Read the issue editorial: Addressing the policy challenges and opportunities of “Big data” by Helen **...**

More

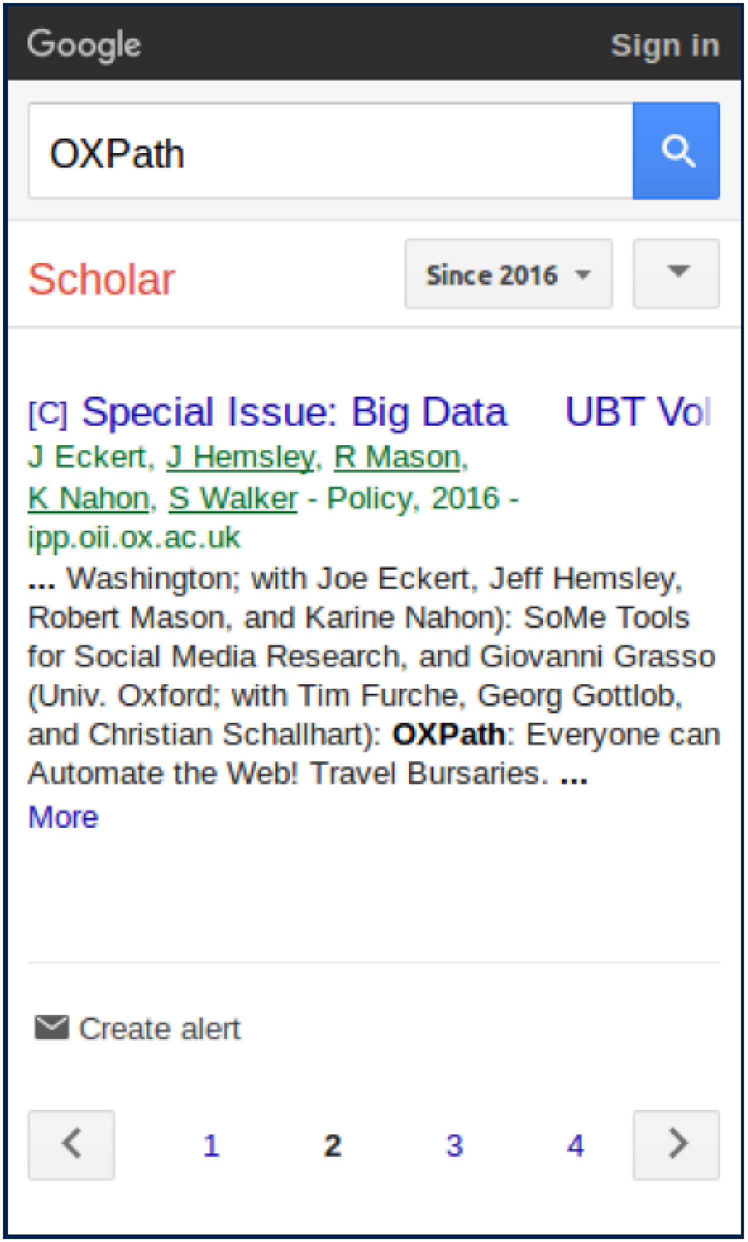
## OXPath Expression

```
doc("https://scholar.google.com")
 //*[@role="search"]//input[@type="text"]/{"OXPath"}
 /../following-sibling::button/{click/}
   //*[@id="gs_ylo_btn"]/{click}
     //following::*[@id="gs_ylo_md"]/a[contains(.,
          "2016")]/{click/}
   /(//*[contains(@class, "next")]/{click/})*
     //div[@class="gs_ri"]//h3/a:<title=string(.)>
```

## XML Output

```
<?xml version="1.0" encoding="UTF-8"?>
<results>
  <title>Tim Furche, Georg Gottlob, [...]</title>
</results>
```

# Example: Navigating Google Scholar



[C] Special Issue: Big Data UBT Vol

J Eckert, J Hemsley, R Mason, K Nahon, S Walker - Policy, 2016 - ipp.oii.ox.ac.uk

... Washington; with Joe Eckert, Jeff Hemsley, Robert Mason, and Karine Nahon): SoMe Tools for Social Media Research, and Giovanni Grasso (Univ. Oxford; with Tim Furche, Georg Gottlob, and Christian Schallhart): **OXPath**: Everyone can Automate the Web! Travel Bursaries. ...

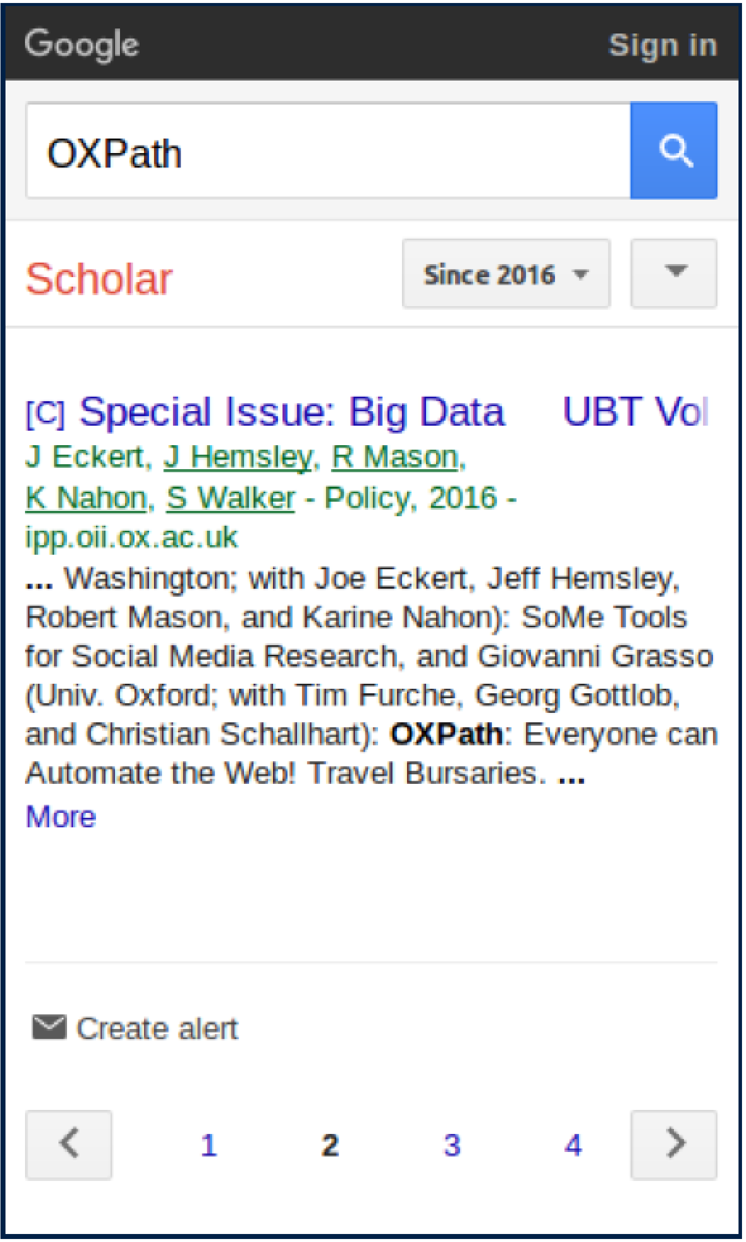
## OXPath Expression

```
doc("https://scholar.google.com")
 //*[@role="search"]//input[@type="text"]/{"OXPath"}
 /../following-sibling::button/{click/}
   //*[@id="gs_ylo_btn"]/{click}
     //following::*[@id="gs_ylo_md"]/a[contains(.,
         "2016")]/{click/}
   /(//*[contains(@class, "next")]/{click/})*
     //div[@class="gs_ri"]//h3/a:<title=string(.)>
```

## XML Output

```
<?xml version="1.0" encoding="UTF-8"?>
<results>
  <title>Tim Furche, Georg Gottlob, [...]</title>
</results>
```

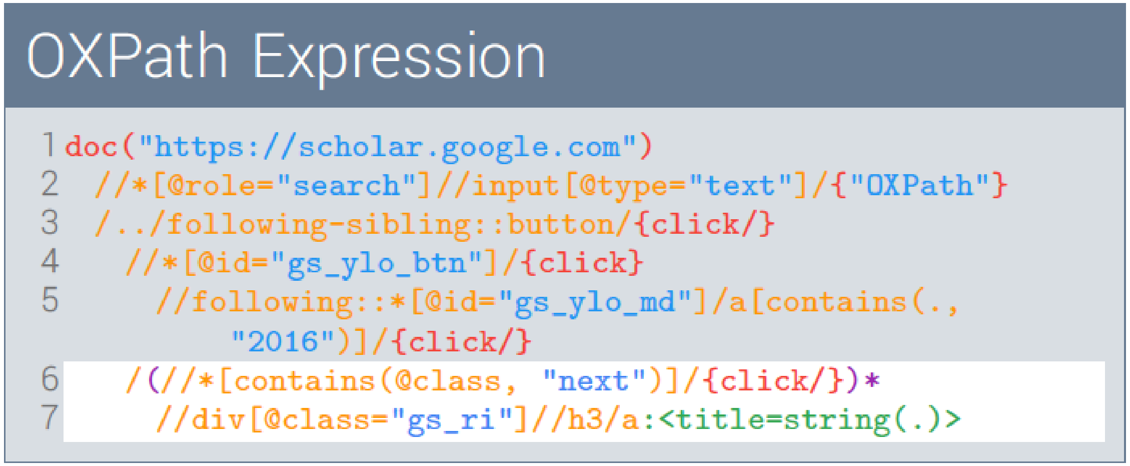
# Example: Navigating Google Scholar



[C] Special Issue: Big Data UBT Vol
J Eckert, J Hemsley, R Mason, K Nahon, S Walker - Policy, 2016 - ipp.oii.ox.ac.uk

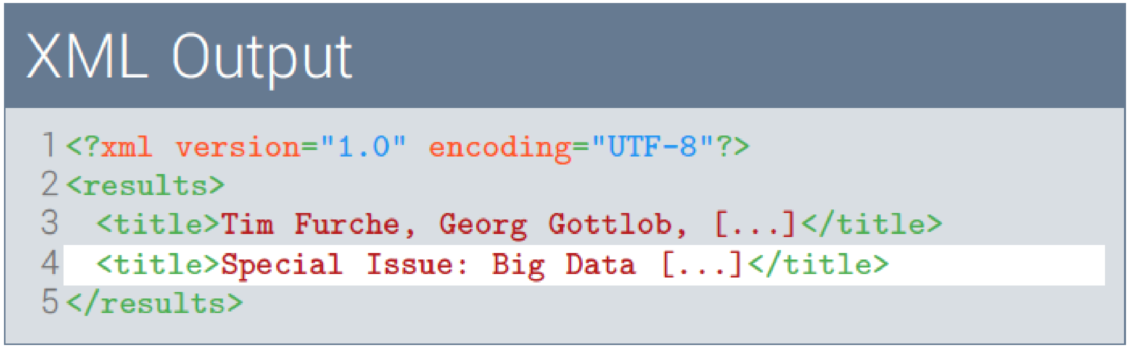
... Washington; with Joe Eckert, Jeff Hemsley, Robert Mason, and Karine Nahon): SoMe Tools for Social Media Research, and Giovanni Grasso (Univ. Oxford; with Tim Furche, Georg Gottlob, and Christian Schallhart): **OXPath**: Everyone can Automate the Web! Travel Bursaries. ...

More



## OXPath Expression

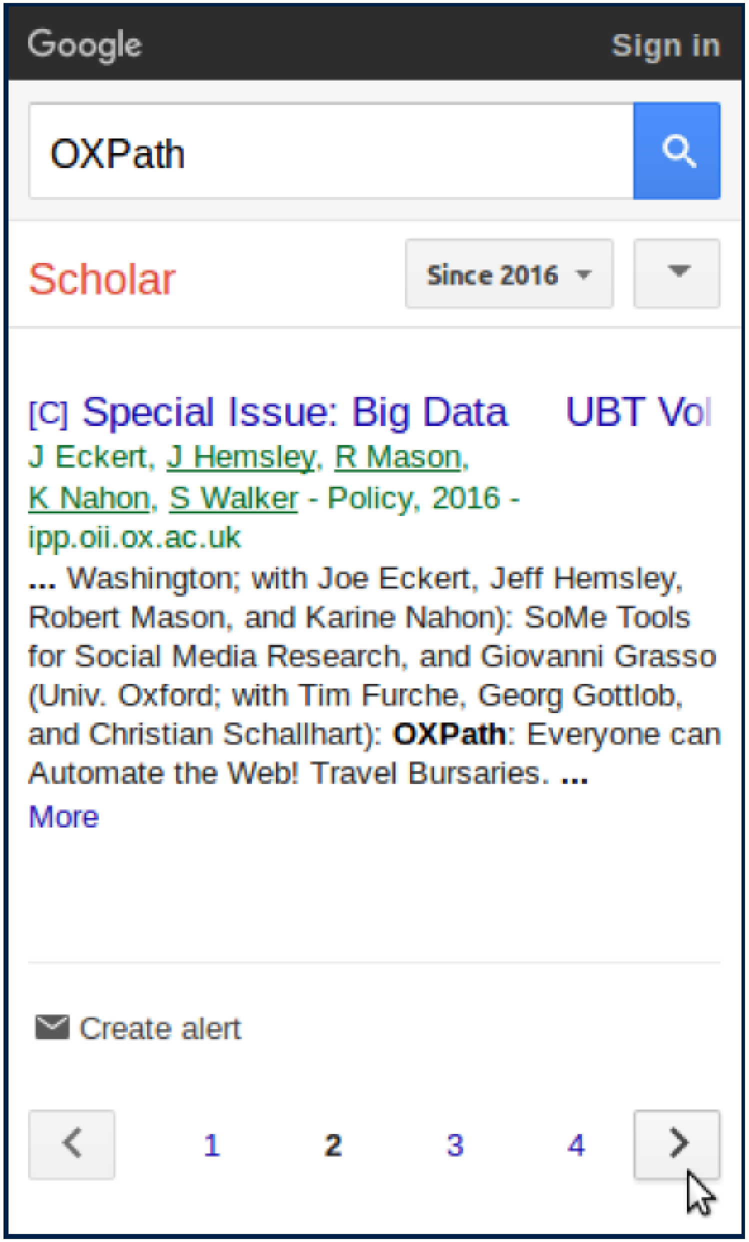
```
doc("https://scholar.google.com")
 //*[@role="search"]//input[@type="text"]/{"OXPath"}
 /../following-sibling::button/{click/}
   //*[@id="gs_ylo_btn"]/{click}
     //following::*[@id="gs_ylo_md"]/a[contains(.,
         "2016")]/{click/}
   /(//*[contains(@class, "next")]/{click/})*
     //div[@class="gs_ri"]//h3/a:<title=string(.)>
```



## XML Output

```
<?xml version="1.0" encoding="UTF-8"?>
<results>
  <title>Tim Furche, Georg Gottlob, [...]</title>
  <title>Special Issue: Big Data [...]</title>
</results>
```

# Example: Navigating Google Scholar



Google Sign in

OXPath

Scholar Since 2016

[C] Special Issue: Big Data UBT Vol

J Eckert, J Hemsley, R Mason, K Nahon, S Walker - Policy, 2016 - ipp.oii.ox.ac.uk

... Washington; with Joe Eckert, Jeff Hemsley, Robert Mason, and Karine Nahon): SoMe Tools for Social Media Research, and Giovanni Grasso (Univ. Oxford; with Tim Furche, Georg Gottlob, and Christian Schallhart): **OXPath**: Everyone can Automate the Web! Travel Bursaries. ...

More

Create alert

< 1 2 3 4 >

## OXPath Expression

```
doc("https://scholar.google.com")
  //*[@role="search"]//input[@type="text"]/{"OXPath"}
  /../following-sibling::button/{click/}
    //*[@id="gs_ylo_btn"]/{click}
      //following::*[@id="gs_ylo_md"]/a[contains(.,
          "2016")]/{click/}
    /(//*[contains(@class, "next")]/{click/})*
      //div[@class="gs_ri"]//h3/a:<title=string(.)>
```

## XML Output

```
<?xml version="1.0" encoding="UTF-8"?>
<results>
  <title>Tim Furche, Georg Gottlob, [...]</title>
  <title>Special Issue: Big Data [...]</title>
  <!--[...]-->
</results>
```

# Advantages of OXPath

- **More powerful** than plain XPath: actions, extraction, transformation, iteration
- Possible to extract from **several pages** in one query
- Somewhat **robust to changes in layout**

**Now in productive use at dblp**

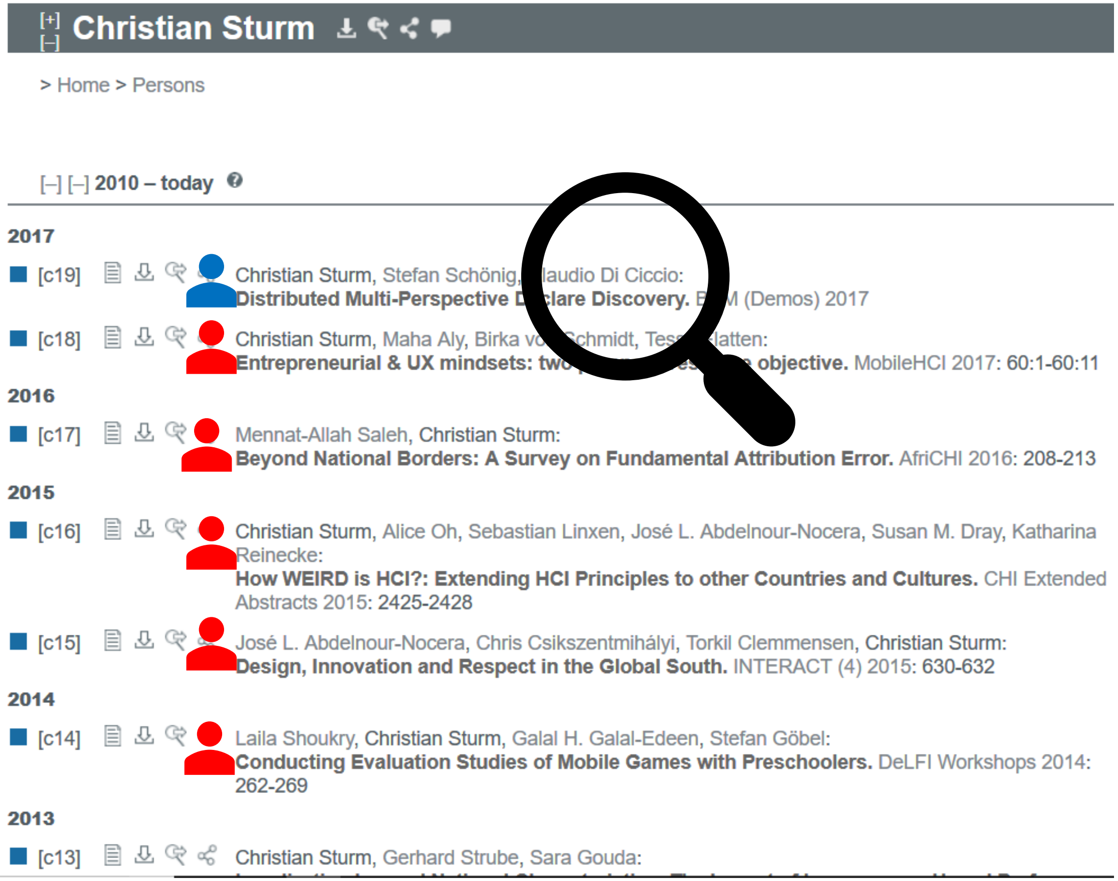
# Outline

- Meta Data Harvesting
- **Author Disambiguation**
- Existing Metadata Collections
- Citations

# Author Disambiguation: Homonyms

dblp
computer science bibliography

**Multiple persons with the same name in the same profile**



[+] [–] **Christian Sturm**

> Home > Persons

[–] [–] **2010 – today**

**2017**

[c19] Christian Sturm, Stefan Schönig, laudio Di Ciccio:
**Distributed Multi-Perspective D clare Discovery.** B M (Demos) 2017

[c18] Christian Sturm, Maha Aly, Birka vo chmidt, Tes latten:
**Entrepreneurial & UX mindsets: two e objective.** MobileHCI 2017: 60:1-60:11

**2016**

[c17] Mennat-Allah Saleh, Christian Sturm:
**Beyond National Borders: A Survey on Fundamental Attribution Error.** AfriCHI 2016: 208-213

**2015**

[c16] Christian Sturm, Alice Oh, Sebastian Linxen, José L. Abdelnour-Nocera, Susan M. Dray, Katharina Reinecke:
**How WEIRD is HCI?: Extending HCI Principles to other Countries and Cultures.** CHI Extended Abstracts 2015: 2425-2428

[c15] José L. Abdelnour-Nocera, Chris Csikszentmihályi, Torkil Clemmensen, Christian Sturm:
**Design, Innovation and Respect in the Global South.** INTERACT (4) 2015: 630-632

**2014**

[c14] Laila Shoukry, Christian Sturm, Galal H. Galal-Edeen, Stefan Göbel:
**Conducting Evaluation Studies of Mobile Games with Preschoolers.** DeLFI Workshops 2014: 262-269

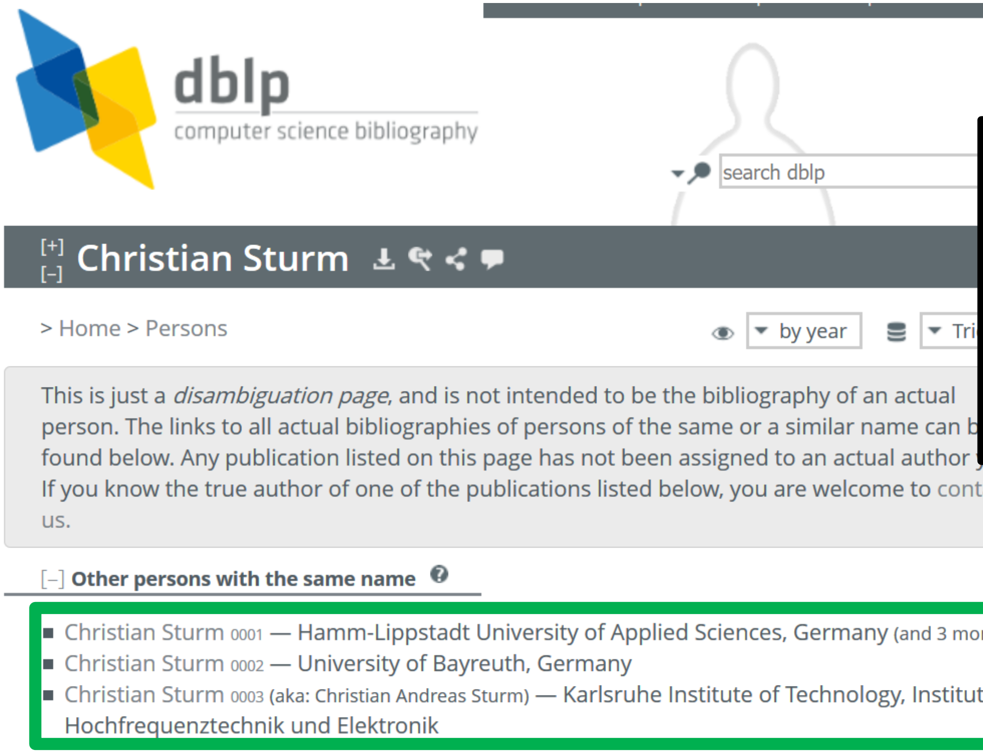
**2013**

[c13] Christian Sturm, Gerhard Strube, Sara Gouda:

**Hard problem for an algorithm (even for a human), may use**

- paper titles/topics
- common coauthors
- publication years
- publication venues
- …

# Author Disambiguation: Homonyms



dblp computer science bibliography

search dblp

## Christian Sturm

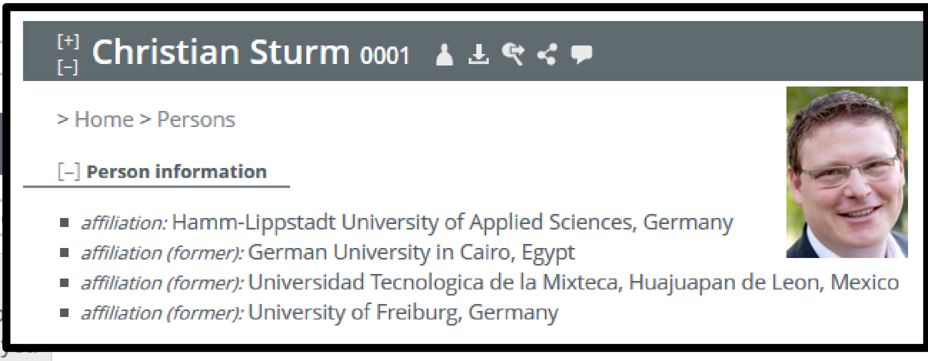
> Home > Persons

by year

This is just a *disambiguation page*, and is not intended to be the bibliography of an actual person. The links to all actual bibliographies of persons of the same or a similar name can b found below. Any publication listed on this page has not been assigned to an actual author y If you know the true author of one of the publications listed below, you are welcome to contact us.

[–] **Other persons with the same name**

- Christian Sturm 0001 — Hamm-Lippstadt University of Applied Sciences, Germany (and 3 more)
- Christian Sturm 0002 — University of Bayreuth, Germany
- Christian Sturm 0003 (aka: Christian Andreas Sturm) — Karlsruhe Institute of Technology, Institut für Hochfrequenztechnik und Elektronik

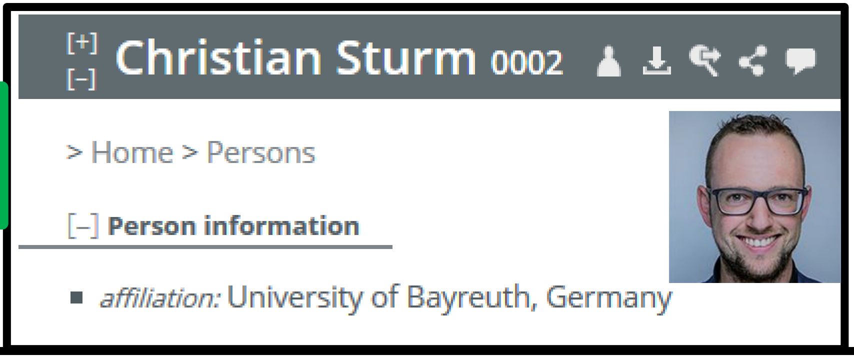


## Christian Sturm 0001

> Home > Persons

[–] **Person information**

- *affiliation:* Hamm-Lippstadt University of Applied Sciences, Germany
- *affiliation (former):* German University in Cairo, Egypt
- *affiliation (former):* Universidad Tecnologica de la Mixteca, Huajuapan de Leon, Mexico
- *affiliation (former):* University of Freiburg, Germany



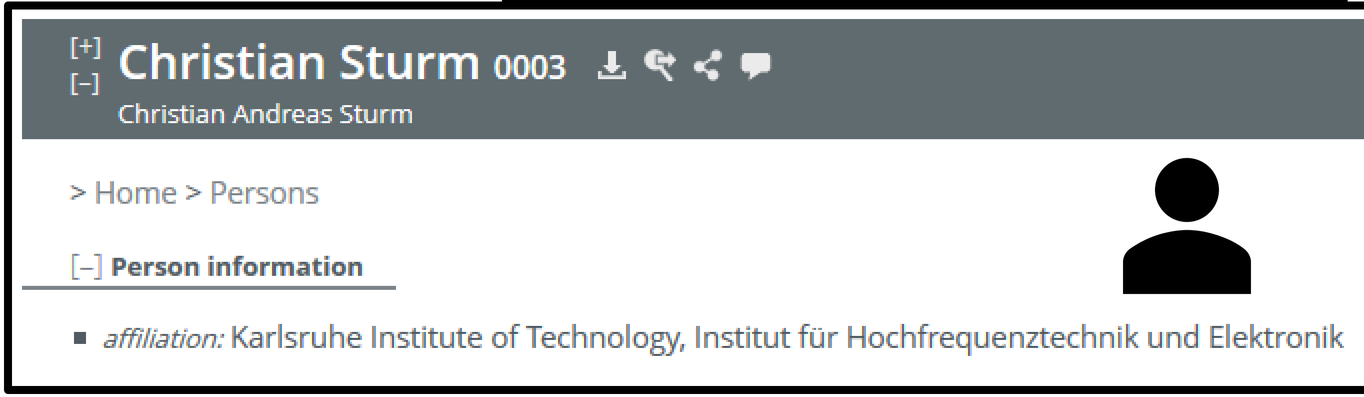
## Christian Sturm 0002

> Home > Persons

[–] **Person information**

- *affiliation:* University of Bayreuth, Germany

**Affiliations would be useful, but usually not available**



## Christian Sturm 0003

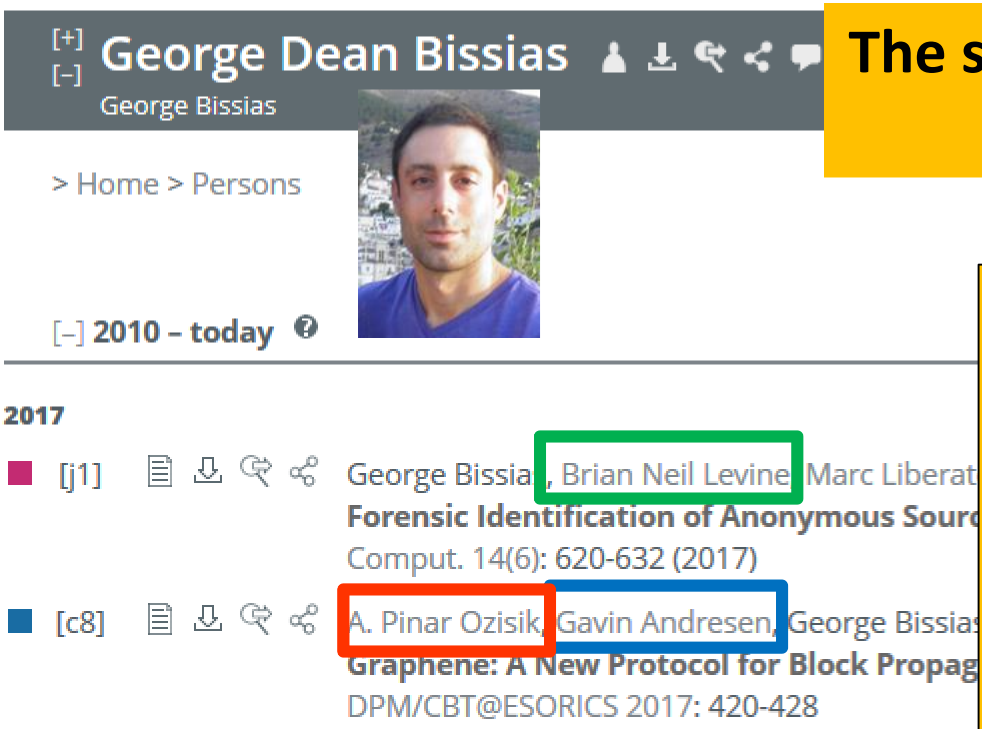
Christian Andreas Sturm

> Home > Persons

[–] **Person information**

- *affiliation:* Karlsruhe Institute of Technology, Institut für Hochfrequenztechnik und Elektronik

# Author Disambiguation: Synonyms



[+]
[–] George Dean Bissias

George Bissias

> Home > Persons

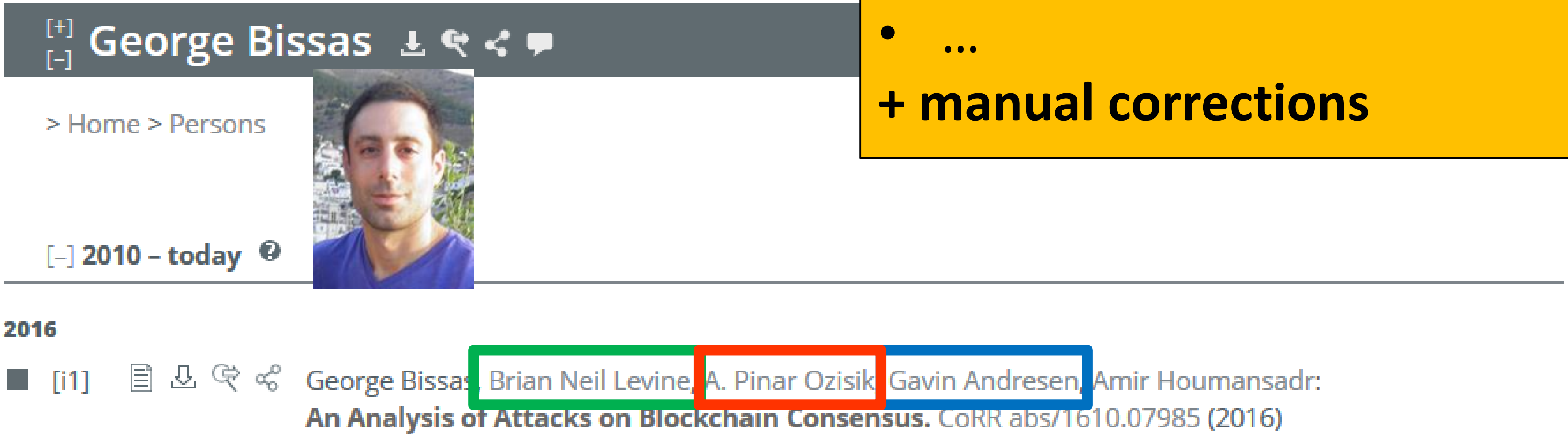
[–] 2010 – today

2017

[j1] George Bissia, Brian Neil Levine, Marc Liberat
**Forensic Identification of Anonymous Sourc**
Comput. 14(6): 620-632 (2017)

[c8] A. Pinar Ozisik, Gavin Andresen, George Bissias
**Graphene: A New Protocol for Block Propag**
DPM/CBT@ESORICS 2017: 420-428

**The same person with different names in different profiles**



[+]
[–] George Bissas

> Home > Persons

[–] 2010 – today

2016

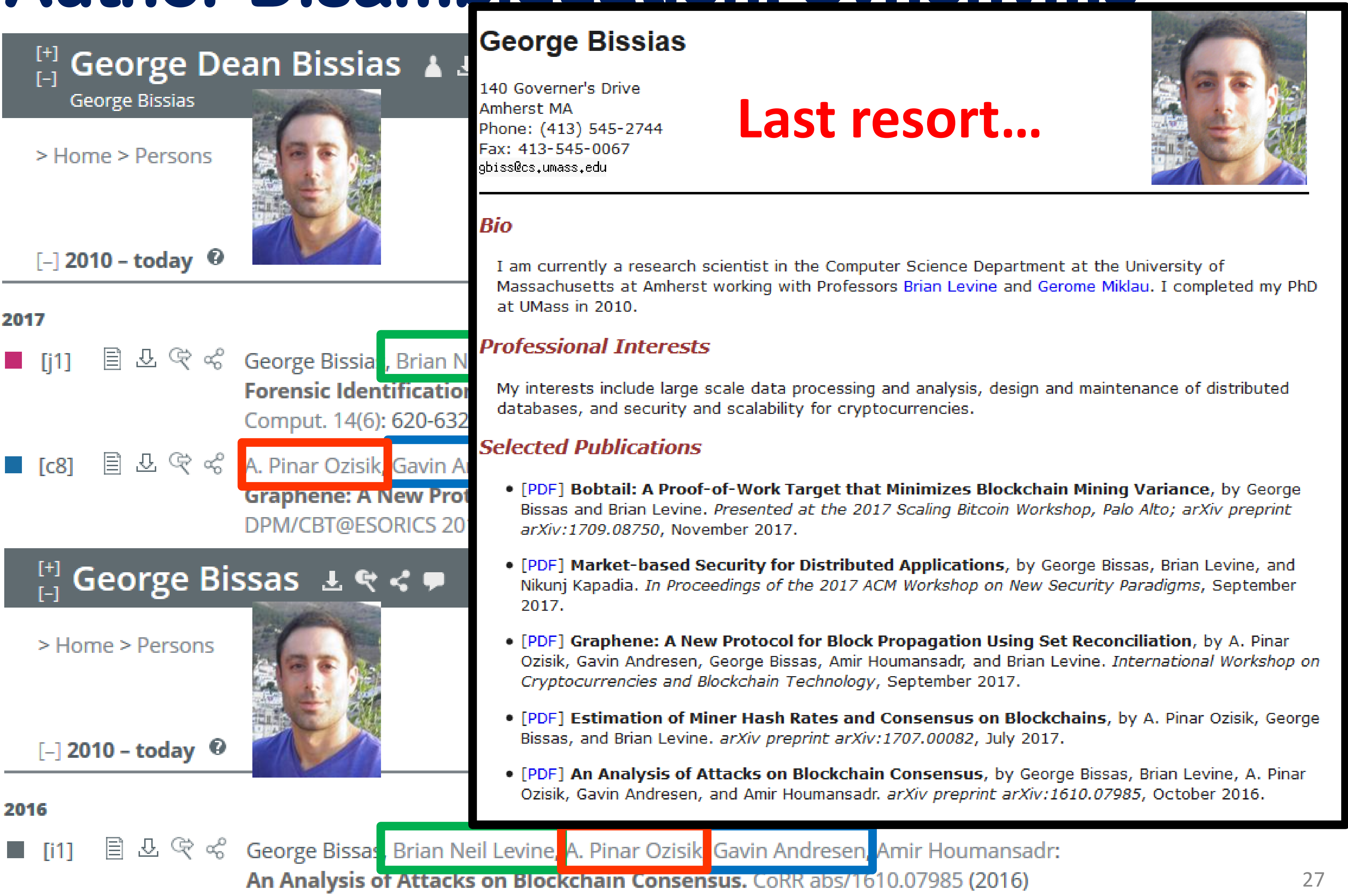
[i1] George Bissas, Brian Neil Levine, A. Pinar Ozisik, Gavin Andresen, Amir Houmansadr:
**An Analysis of Attacks on Blockchain Consensus.** CoRR abs/1610.07985 (2016)

**Identify pairs of candidate profiles such that**

- Small name difference
- Common coauthors
- Common venues
- Common topics
- …

**+ manual corrections**

# Author Disambiguation: Synonyms



[+] [–] George Dean Bissias

George Bissias

> Home > Persons

[–] 2010 – today

2017

[j1] George Bissia, Brian N
Forensic Identification
Comput. 14(6): 620-632

[c8] A. Pinar Ozisik, Gavin A
Graphene: A New Prot
DPM/CBT@ESORICS 20

[+] [–] George Bissas

> Home > Persons

[–] 2010 – today

2016

[i1] George Bissas, Brian Neil Levine, A. Pinar Ozisik, Gavin Andresen, Amir Houmansadr:
**An Analysis of Attacks on Blockchain Consensus.** CoRR abs/1610.07985 (2016)

## George Bissias

140 Governer's Drive
Amherst MA
Phone: (413) 545-2744
Fax: 413-545-0067
gbiss@cs.umass.edu

Last resort...

### *Bio*

I am currently a research scientist in the Computer Science Department at the University of Massachusetts at Amherst working with Professors Brian Levine and Gerome Miklau. I completed my PhD at UMass in 2010.

### *Professional Interests*

My interests include large scale data processing and analysis, design and maintenance of distributed databases, and security and scalability for cryptocurrencies.

### *Selected Publications*

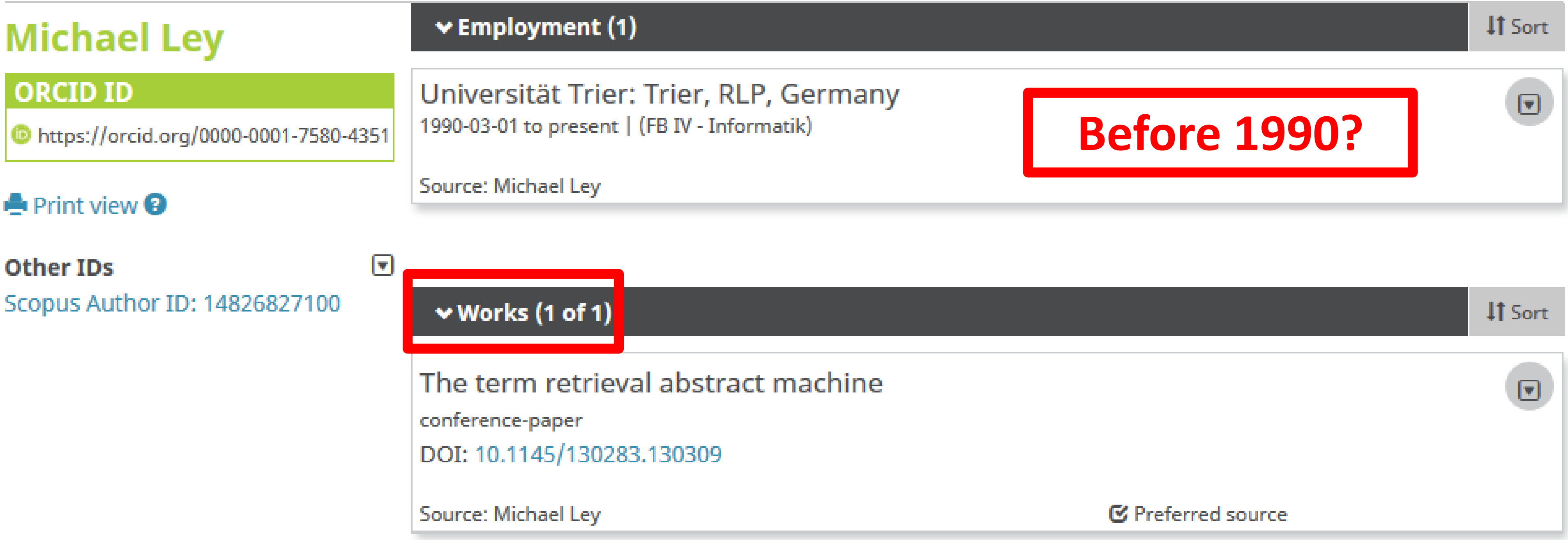
- [PDF] **Bobtail: A Proof-of-Work Target that Minimizes Blockchain Mining Variance**, by George Bissas and Brian Levine. *Presented at the 2017 Scaling Bitcoin Workshop, Palo Alto; arXiv preprint arXiv:1709.08750*, November 2017.
- [PDF] **Market-based Security for Distributed Applications**, by George Bissas, Brian Levine, and Nikunj Kapadia. *In Proceedings of the 2017 ACM Workshop on New Security Paradigms*, September 2017.
- [PDF] **Graphene: A New Protocol for Block Propagation Using Set Reconciliation**, by A. Pinar Ozisik, Gavin Andresen, George Bissas, Amir Houmansadr, and Brian Levine. *International Workshop on Cryptocurrencies and Blockchain Technology*, September 2017.
- [PDF] **Estimation of Miner Hash Rates and Consensus on Blockchains**, by A. Pinar Ozisik, George Bissas, and Brian Levine. *arXiv preprint arXiv:1707.00082*, July 2017.
- [PDF] **An Analysis of Attacks on Blockchain Consensus**, by George Bissas, Brian Levine, A. Pinar Ozisik, Gavin Andresen, and Amir Houmansadr. *arXiv preprint arXiv:1610.07985*, October 2016.

**Observation:**
Additional meta data can improve the quality of the detection of synonyms and homonyms.

# Example: ORCID

> 4.3 million ORCIDs

- Provides **persistent digital identifier** for authors
- Includes additional **author-provided meta data** about publications, affiliations, …
- API & dumps



Michael Ley

ORCID ID

https://orcid.org/0000-0001-7580-4351

Print view

Other IDs

Scopus Author ID: 14826827100

Employment (1) | Sort

Universität Trier: Trier, RLP, Germany

1990-03-01 to present | (FB IV - Informatik)

Source: Michael Ley

Works (1 of 1) | Sort

The term retrieval abstract machine

conference-paper

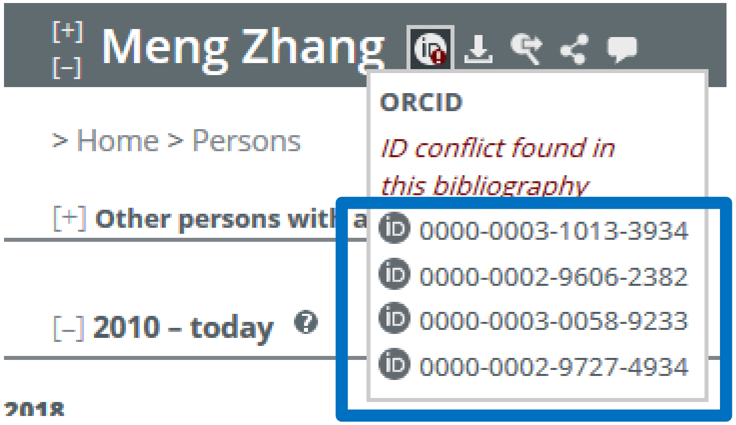
DOI: 10.1145/130283.130309

Source: Michael Ley ☑ Preferred source

Before 1990?

**Data often incomplete or not fully correct**

# ORCID for Homonym Detection



[+]
[–] Meng Zhang

> Home > Persons

[+] Other persons with a

[–] 2010 – today

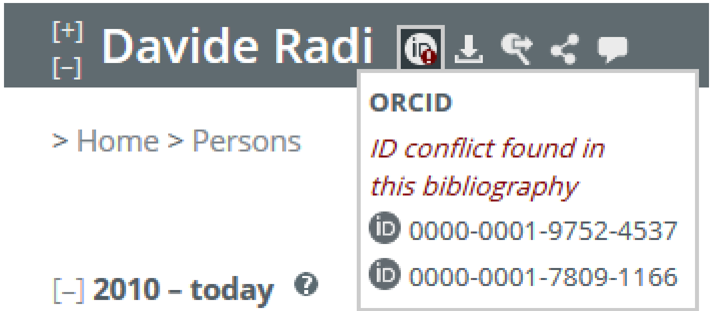
2018

ORCID

*ID conflict found in this bibliography*

0000-0003-1013-3934
0000-0002-9606-2382
0000-0003-0058-9233
0000-0002-9727-4934

ORCIDs of authors with this name who claimed at least one publication in this profile



[+]
[–] Davide Radi

> Home > Persons

[–] 2010 – today

ORCID

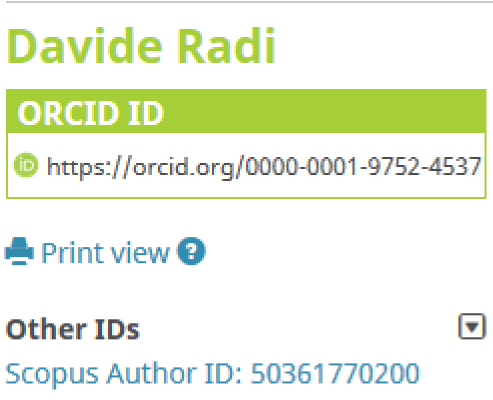
*ID conflict found in this bibliography*

0000-0001-9752-4537
0000-0001-7809-1166

**After import of 625,000 ORCIDS:**
**1,000 candidates for homonyms**

**Top candidate:**
**10 persons in one profile**

**BUT:**



Davide Radi

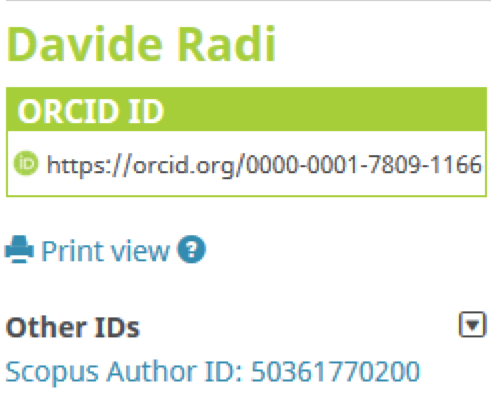
ORCID ID

https://orcid.org/0000-0001-9752-4537

Print view

Other IDs

Scopus Author ID: 50361770200



Davide Radi

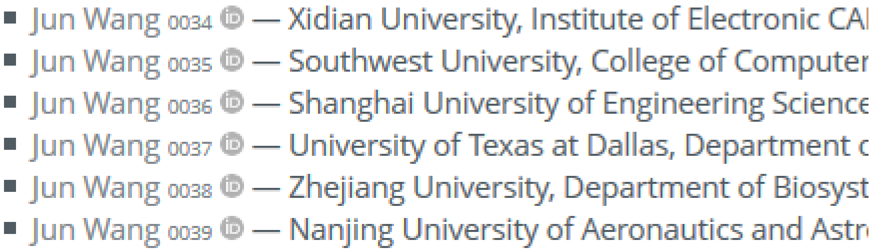
ORCID ID

https://orcid.org/0000-0001-7809-1166

Print view

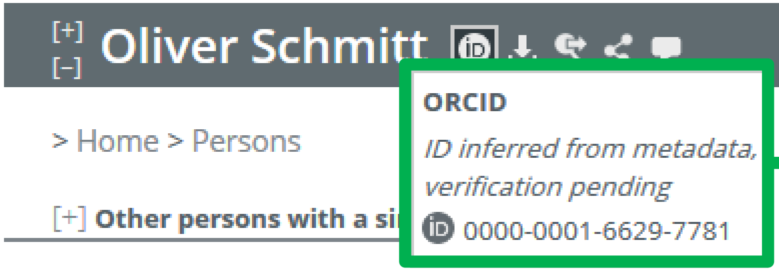
Other IDs

Scopus Author ID: 50361770200



- Jun Wang 0034 — Xidian University, Institute of Electronic CAI
- Jun Wang 0035 — Southwest University, College of Computer
- Jun Wang 0036 — Shanghai University of Engineering Science
- Jun Wang 0037 — University of Texas at Dallas, Department c
- Jun Wang 0038 — Zhejiang University, Department of Biosyst
- Jun Wang 0039 — Nanjing University of Aeronautics and Astr

# ORCID for Synonym Detection



[+] [–] Oliver Schmitt

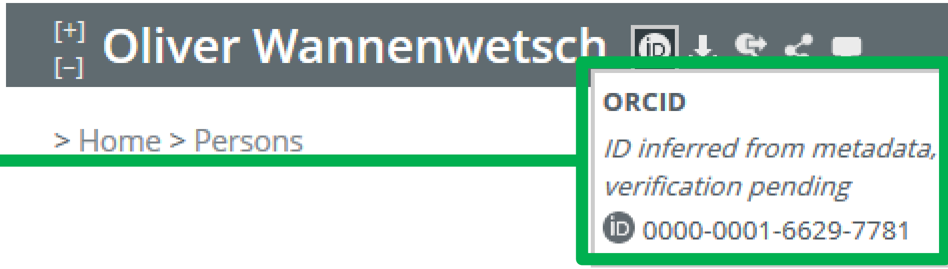
> Home > Persons

[+] Other persons with a si

ORCID

*ID inferred from metadata, verification pending*

0000-0001-6629-7781



[+] [–] Oliver Wannenwetsch

> Home > Persons

ORCID

*ID inferred from metadata, verification pending*

0000-0001-6629-7781

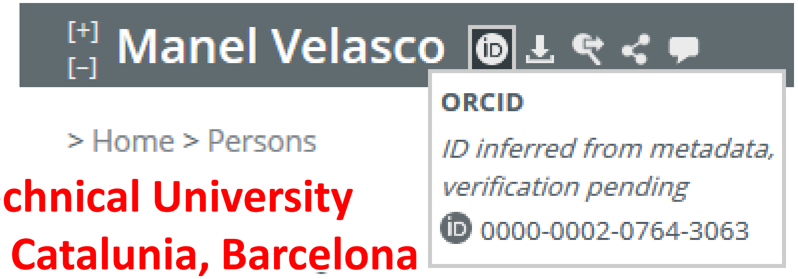
**Profiles with common ORCID include papers from the same author (but maybe other papers as well due to homonyms)**

**After import of 625,000 ORCIDS: 4,500 candidates for synonyms**

**Top candidate: 6 profiles with same ORCID**

**X. Xu, X. W. Xu, X. William Xu, Xun Xu, Xun W. Xu, Xun William Xu**

**BUT:**



[+] [–] Manel Velasco

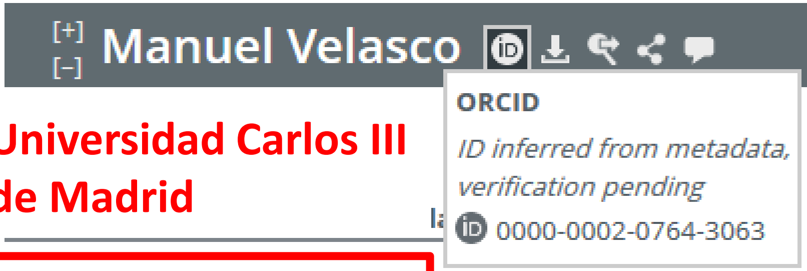
> Home > Persons

ORCID

*ID inferred from metadata, verification pending*

0000-0002-0764-3063

**Technical University of Catalunia, Barcelona**



[+] [–] Manuel Velasco

ORCID

*ID inferred from metadata, verification pending*

0000-0002-0764-3063

**Universidad Carlos III de Madrid**

Source: Scopus to ORCID

# Outline

- Meta Data Harvesting
- Author Disambiguation
- **Existing Metadata Collections**
- Citations

# Useful information not (always) in dblp

- Author affiliations → better disambiguation, better search
- Keywords
- Topics
- Abstracts
- Full texts

better search

- Incoming and outgoing citations
- Performance indicators

better result ranking
better conference selection

- …

# Sources for Bibliographic Metadata

- Dblp.org
- Semantic Scholar
- Aminer Open academic graph (includes Microsoft Academic Graph)
- Springer SciGraph
- CrossRef
- OpenCitations
- …

# Overview: properties of sources

| | Semantic Scholar | OAG - Aminer | OAG – Microsoft Academic | Springer SciGraph | CrossRef | Open Citations |
|---|---|---|---|---|---|---|
| coverage | CS | universal | universal | Springer | universal | universal |
| # publs | 7.2 million | 154 million | 166 million | ~12 million | 96 million | ~300,000 |
| in dblp | 1.45 million | 3.46 million | 3.57 million | ? | ? | ? |
| access | dump | dump | dump | API, dump | API | API, dump |
| size | 20 GB | 39 GB | 103 GB | ~200 GB | - | 3.5 GB |
| date | Oct 2017 | Mar 2017 | Jun 2017 | Nov 2017 | live | Dec 2017 |
| Keywords | 🟩 | 🟩 | 🟩 | 🟥 | 🟥 | |
| Topics | 🟥 | 🟥 | 🟩 | 🟩 | 🟥 | |
| Abstracts | 🟩 | 🟩 | 🟩 | 🟩 | partial | |
| Full-texts | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 |
| Citations | 🟩 | 🟩 | 🟩 | planned | partial | 🟩 |
| DOIs | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 |
| Author aff. | email | 🟩 | 🟩 | 🟩 | partial | |
| Funding | 🟥 | 🟥 | 🟥 | 🟩 | partial | |

# SpringerNature SciGraph

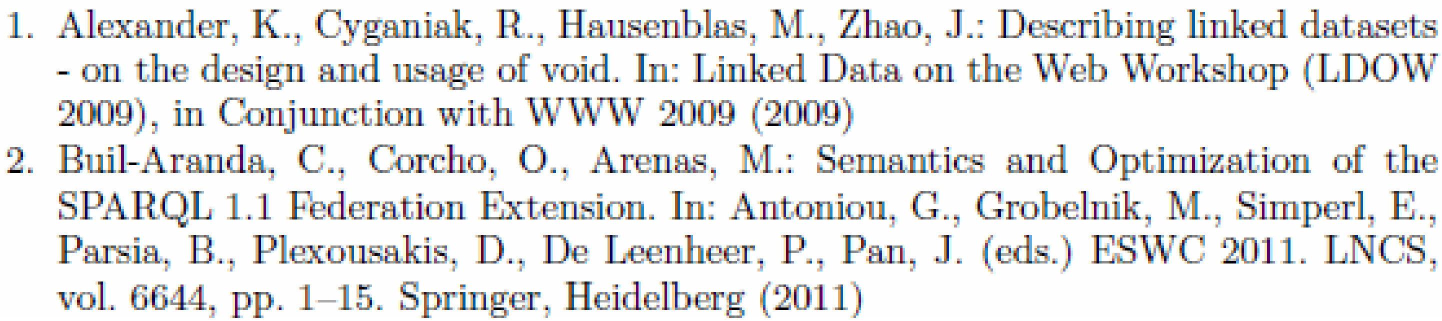
- Linked Open Data with rich ontology
- **funders**, **research projects**, **conferences**, **affiliations** and **publications** from SpringerNature and partners
- extension to **citations**, **patents**, **clinical trials** and **usage numbers** planned
- **CC BY 4.0 license** (NC for abstracts)

**http://www.springernature.com/gp/researchers/scigraph**

# OpenCitations

- **Initiative for Open Citations (I4OC)**:
  collaboration between scholarly publishers and researchers to promote the **unrestricted availability of scholarly citation data**
- As of January 2018, **50% of publications at CrossRef** with open references
- **OpenCitations**:
  publishes open citations from CrossRef as RDF-based collection, using SPAR ontology

# CrossRef Example



1. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing linked datasets - on the design and usage of void. In: Linked Data on the Web Workshop (LDOW 2009), in Conjunction with WWW 2009 (2009)
2. Buil-Aranda, C., Corcho, O., Arenas, M.: Semantics and Optimization of the SPARQL 1.1 Federation Extension. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.) ESWC 2011. LNCS, vol. 6644, pp. 1–15. Springer, Heidelberg (2011)

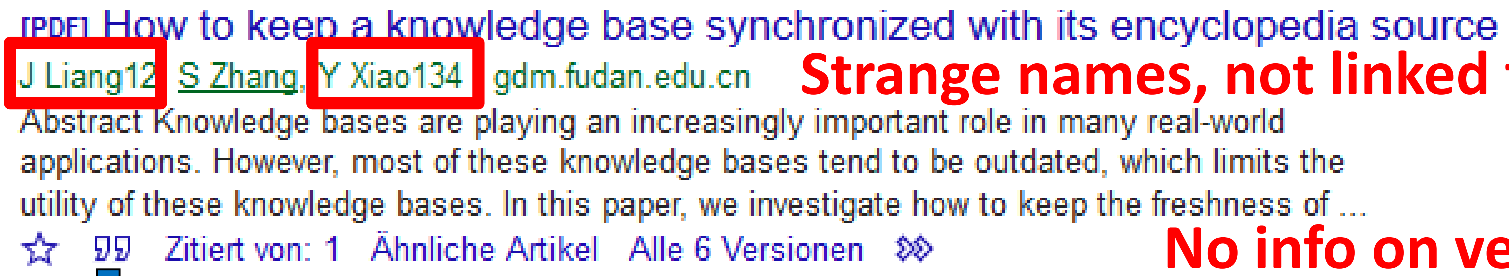
```
"reference":[
{"key":"38_CR1","unstructured":"Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing linked datasets - on the design and usage of void. In: Linked Data on the Web Workshop (LDOW 2009), in Conjunction with WWW 2009 (2009)"},
{"key":"38_CR2","unstructured":"Buil-Aranda, C., Corcho, O., Arenas, M.: Semantics and Optimization of the SPARQL 1.1 Federation Extension. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.) ESWC 2011. LNCS, vol.\u00a06644, pp. 1\u201315. Springer, Heidelberg (2011)","DOI":"10.1007\/978-3-642-21064-8_1","doi-asserted-by":"crossref"},
…]
```

http://api.crossref.org/works/10.1007/978-3-642-25073-6_38

# Problems of these Collections

- Update Frequency
- Data Quality
- Completeness / Coverage / Sparsity

# Data Quality: Automatic Extraction



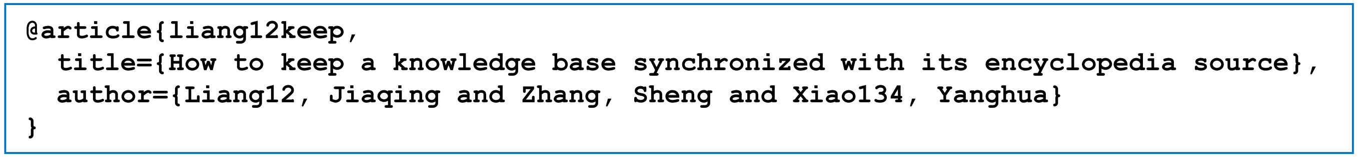
[PDF] How to keep a knowledge base synchronized with its encyclopedia source

J Liang12, S Zhang, Y Xiao134 - gdm.fudan.edu.cn

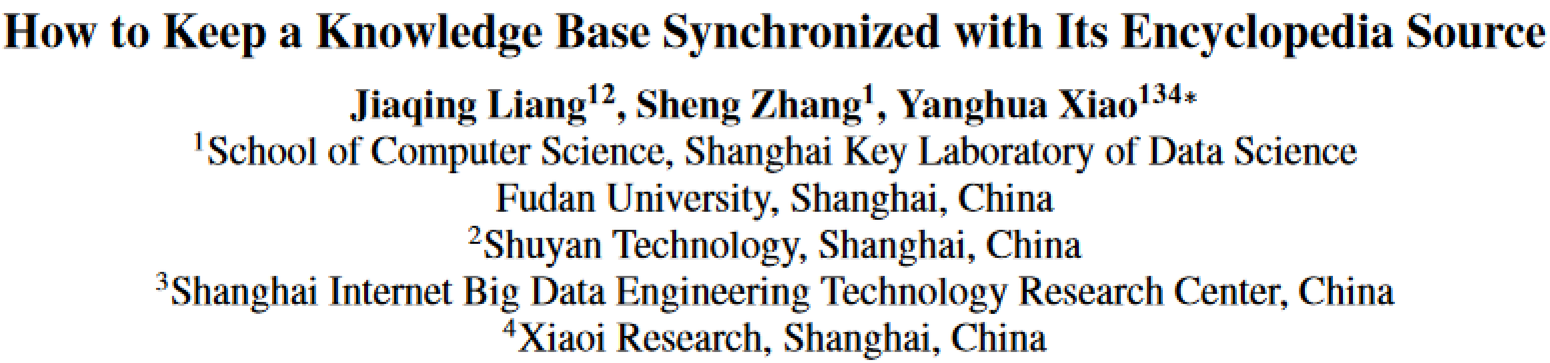
Abstract Knowledge bases are playing an increasingly important role in many real-world applications. However, most of these knowledge bases tend to be outdated, which limits the utility of these knowledge bases. In this paper, we investigate how to keep the freshness of ...

Strange names, not linked to a profile

No info on venue, year ,...

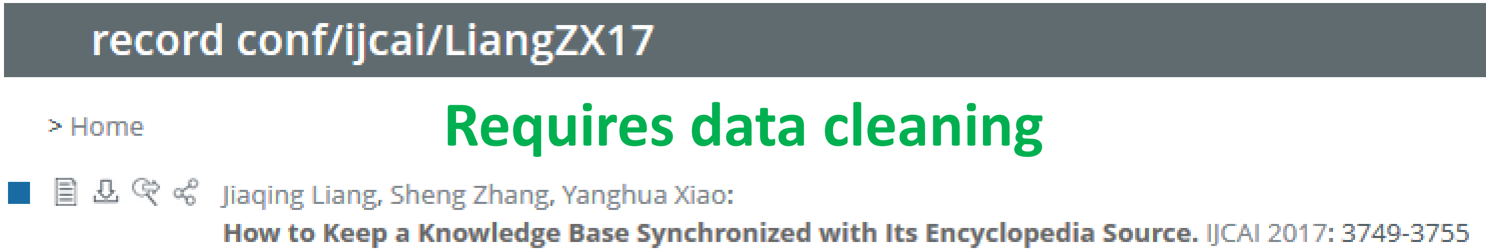


```
@article{liang12keep,
  title={How to keep a knowledge base synchronized with its encyclopedia source},
  author={Liang12, Jiaqing and Zhang, Sheng and Xiao134, Yanghua}
}
```



How to Keep a Knowledge Base Synchronized with Its Encyclopedia Source

Jiaqing Liang[12], Sheng Zhang[1], Yanghua Xiao[134*]
[1]School of Computer Science, Shanghai Key Laboratory of Data Science
Fudan University, Shanghai, China
[2]Shuyan Technology, Shanghai, China
[3]Shanghai Internet Big Data Engineering Technology Research Center, China
[4]Xiaoi Research, Shanghai, China



record conf/ijcai/LiangZX17

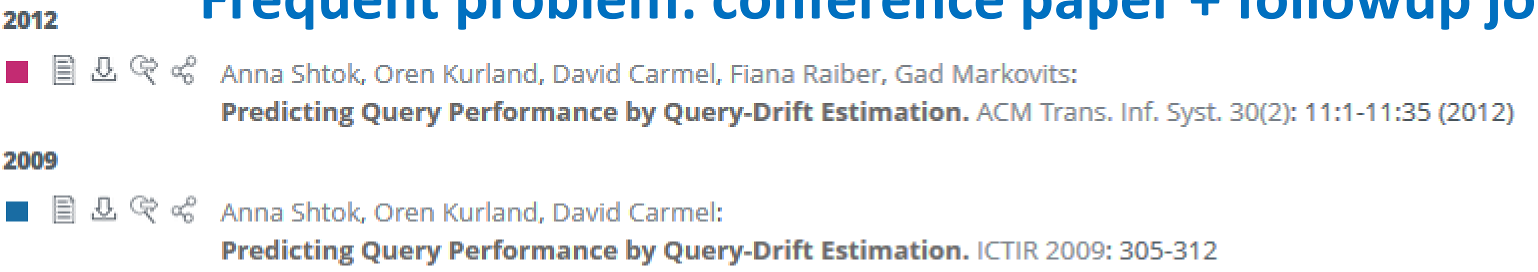
> Home

Jiaqing Liang, Sheng Zhang, Yanghua Xiao:
How to Keep a Knowledge Base Synchronized with Its Encyclopedia Source. IJCAI 2017: 3749-3755

Requires data cleaning

# Data Quality: What is a Publication?

## Frequent problem: conference paper + followup journal paper

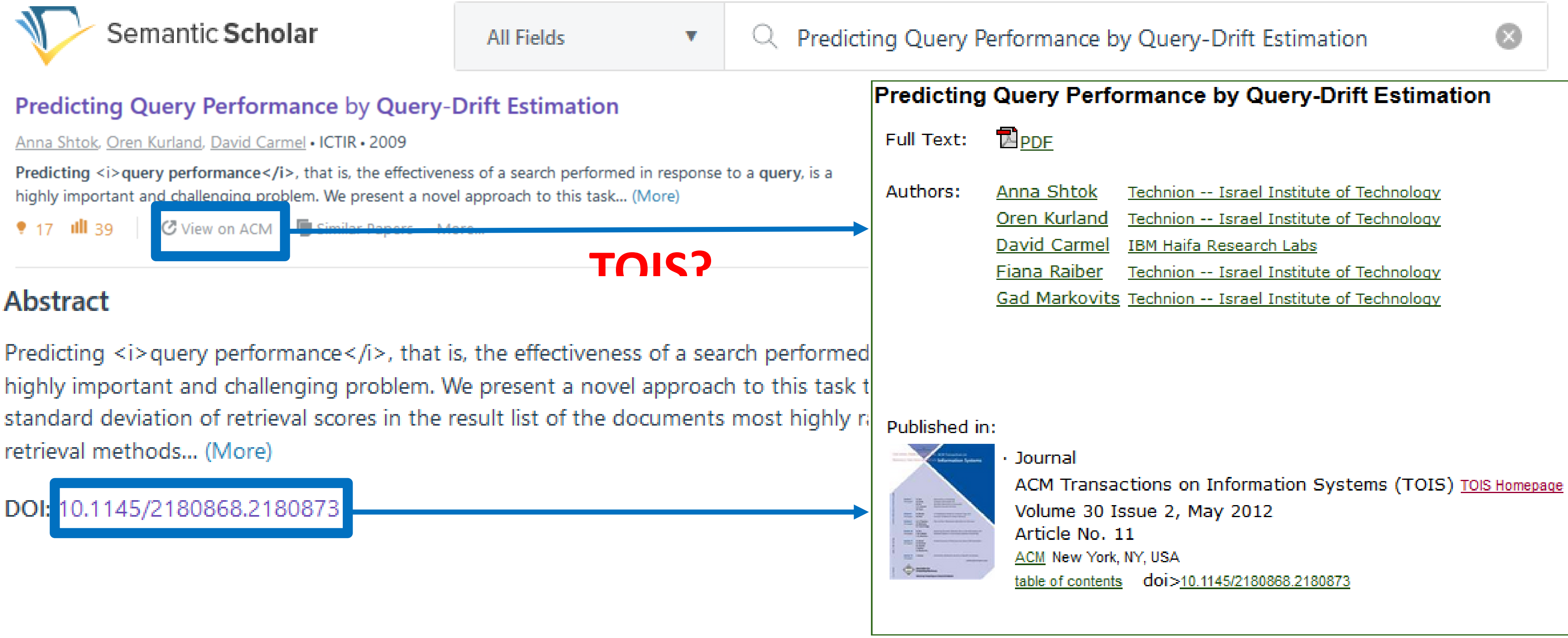


2012

🟥 Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber, Gad Markovits:
**Predicting Query Performance by Query-Drift Estimation.** ACM Trans. Inf. Syst. 30(2): 11:1-11:35 (2012)

2009

🟦 Anna Shtok, Oren Kurland, David Carmel:
**Predicting Query Performance by Query-Drift Estimation.** ICTIR 2009: 305-312



Predicting Query Performance by Query-Drift Estimation

Anna Shtok, Oren Kurland, David Carmel • ICTIR • 2009

Predicting <i>query performance</i>, that is, the effectiveness of a search performed in response to a query, is a highly important and challenging problem. We present a novel approach to this task... (More)

Abstract

Predicting <i>query performance</i>, that is, the effectiveness of a search performed highly important and challenging problem. We present a novel approach to this task t standard deviation of retrieval scores in the result list of the documents most highly r retrieval methods... (More)

DOI: 10.1145/2180868.2180873

**Predicting Query Performance by Query-Drift Estimation**

Full Text: PDF

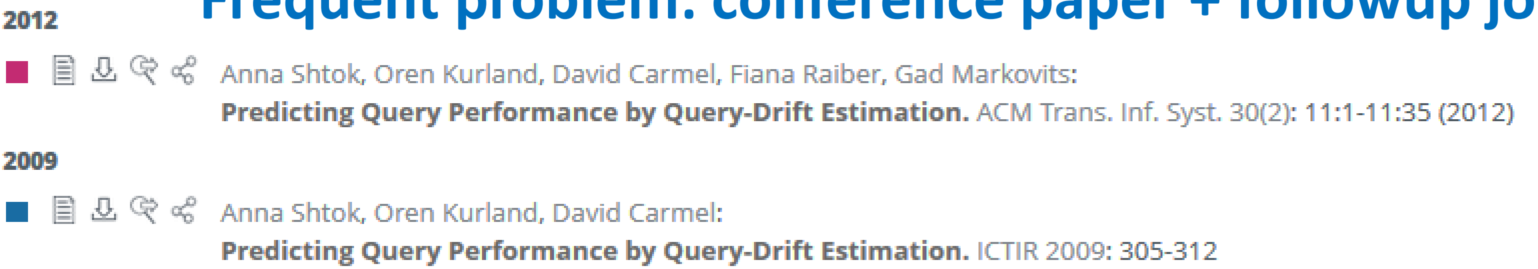
| Authors: | | |
|---|---|---|
| | Anna Shtok | Technion -- Israel Institute of Technology |
| | Oren Kurland | Technion -- Israel Institute of Technology |
| | David Carmel | IBM Haifa Research Labs |
| | Fiana Raiber | Technion -- Israel Institute of Technology |
| | Gad Markovits | Technion -- Israel Institute of Technology |

Published in:

- Journal
  ACM Transactions on Information Systems (TOIS) TOIS Homepage
  Volume 30 Issue 2, May 2012
  Article No. 11
  ACM New York, NY, USA
  table of contents doi>10.1145/2180868.2180873

# Data Quality: What is a Publication?

## Frequent problem: conference paper + followup journal paper

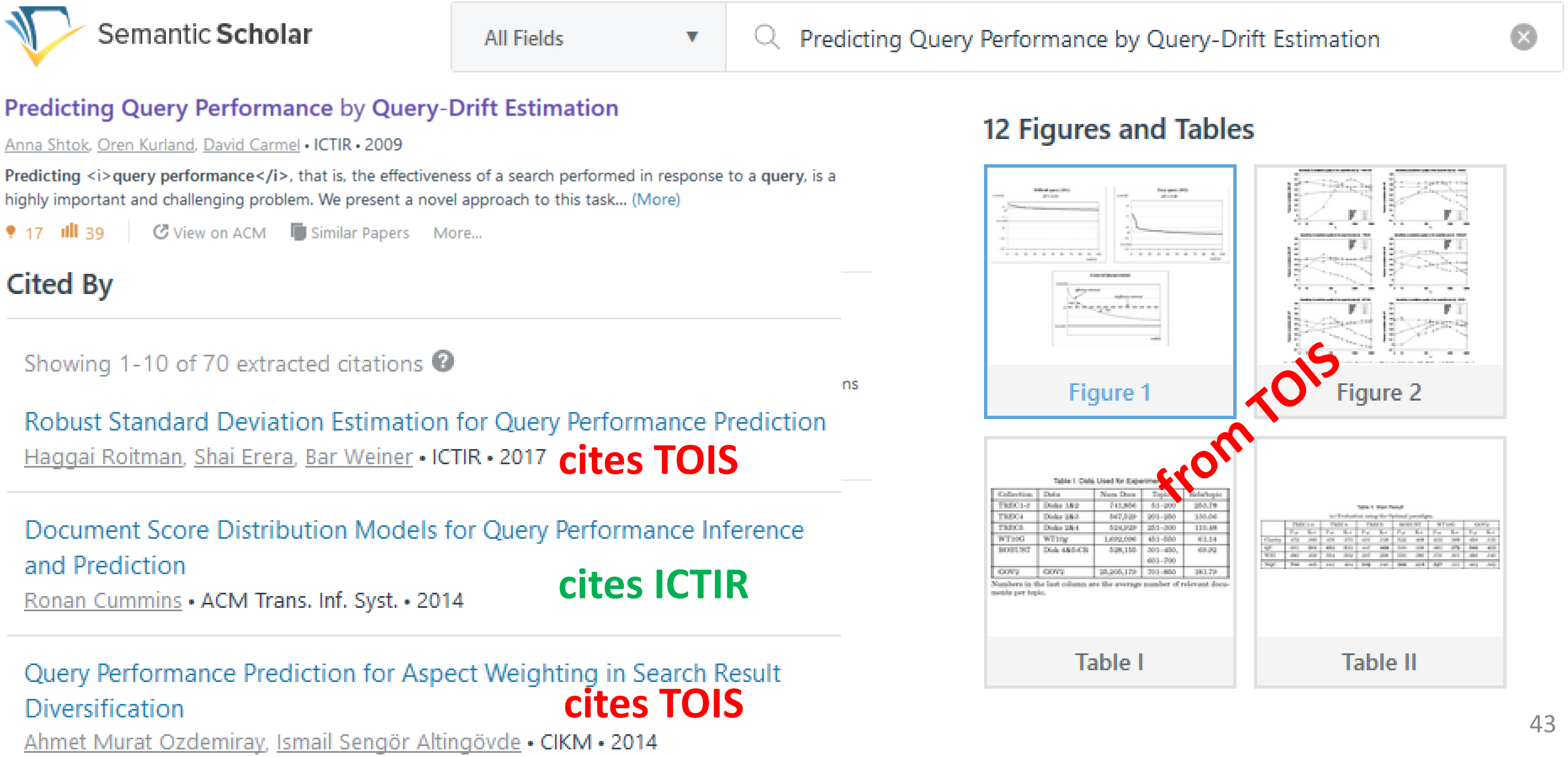


2012

Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber, Gad Markovits:
**Predicting Query Performance by Query-Drift Estimation.** ACM Trans. Inf. Syst. 30(2): 11:1-11:35 (2012)

2009

Anna Shtok, Oren Kurland, David Carmel:
**Predicting Query Performance by Query-Drift Estimation.** ICTIR 2009: 305-312



Predicting Query Performance by Query-Drift Estimation

Anna Shtok, Oren Kurland, David Carmel • ICTIR • 2009

**Predicting** <i>**query performance**</i>, that is, the effectiveness of a search performed in response to a **query**, is a highly important and challenging problem. We present a novel approach to this task... (More)

Cited By

Showing 1-10 of 70 extracted citations

ns

Robust Standard Deviation Estimation for Query Performance Prediction
Haggai Roitman, Shai Erera, Bar Weiner • ICTIR • 2017 **cites TOIS**

Document Score Distribution Models for Query Performance Inference and Prediction
Ronan Cummins • ACM Trans. Inf. Syst. • 2014 **cites ICTIR**

Query Performance Prediction for Aspect Weighting in Search Result Diversification
Ahmet Murat Ozdemiray, Ismail Sengör Altingövde • CIKM • 2014 **cites TOIS**

12 Figures and Tables

Figure 1

Figure 2

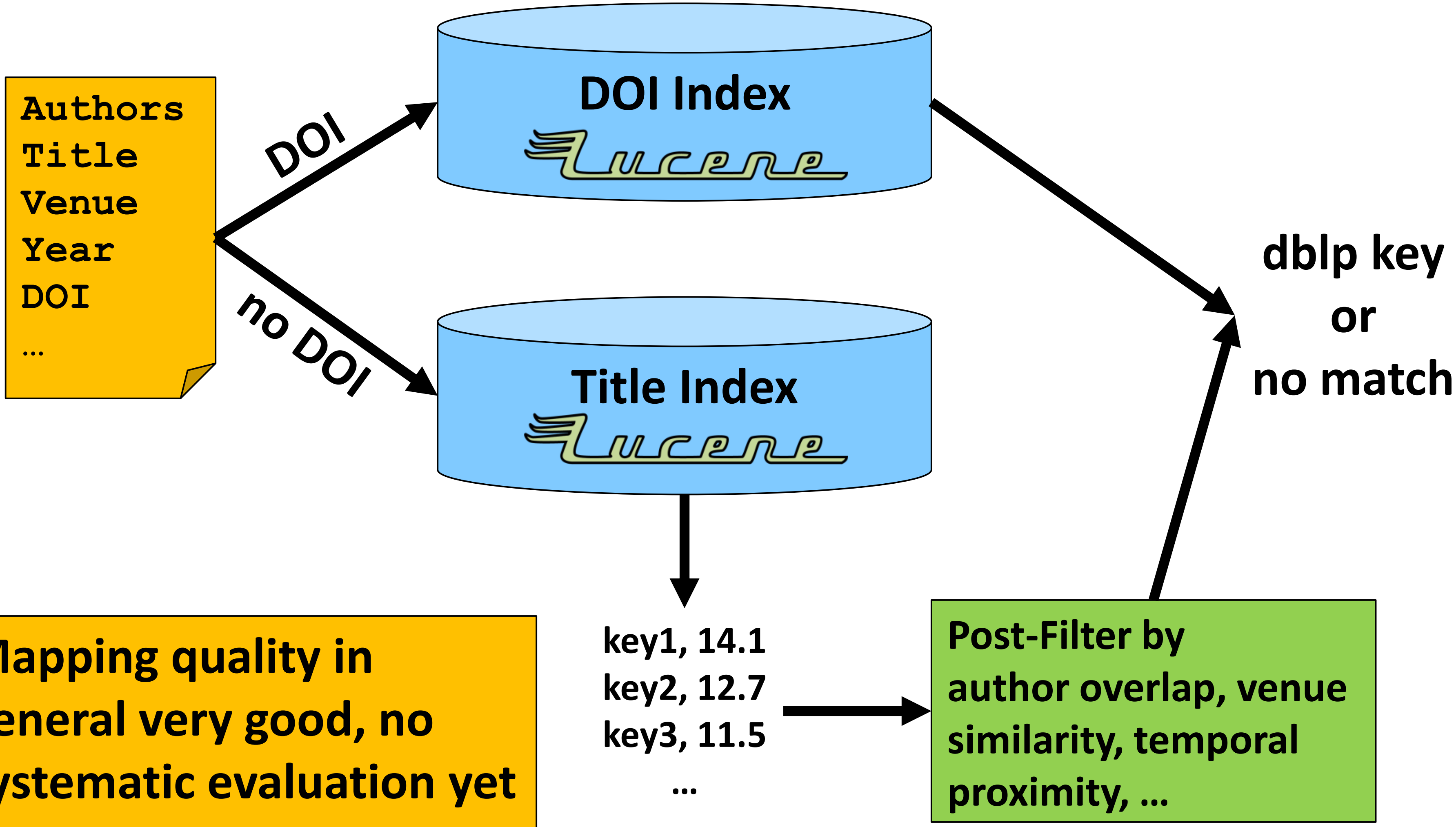
Table I

Table II

**from TOIS**

# Towards Quantifying Coverage: Mapping papers to dblp

**Preprocessing**: Index all dblp entries in Lucene



Authors
Title
Venue
Year
DOI
…

key1, 14.1
key2, 12.7
key3, 11.5
…

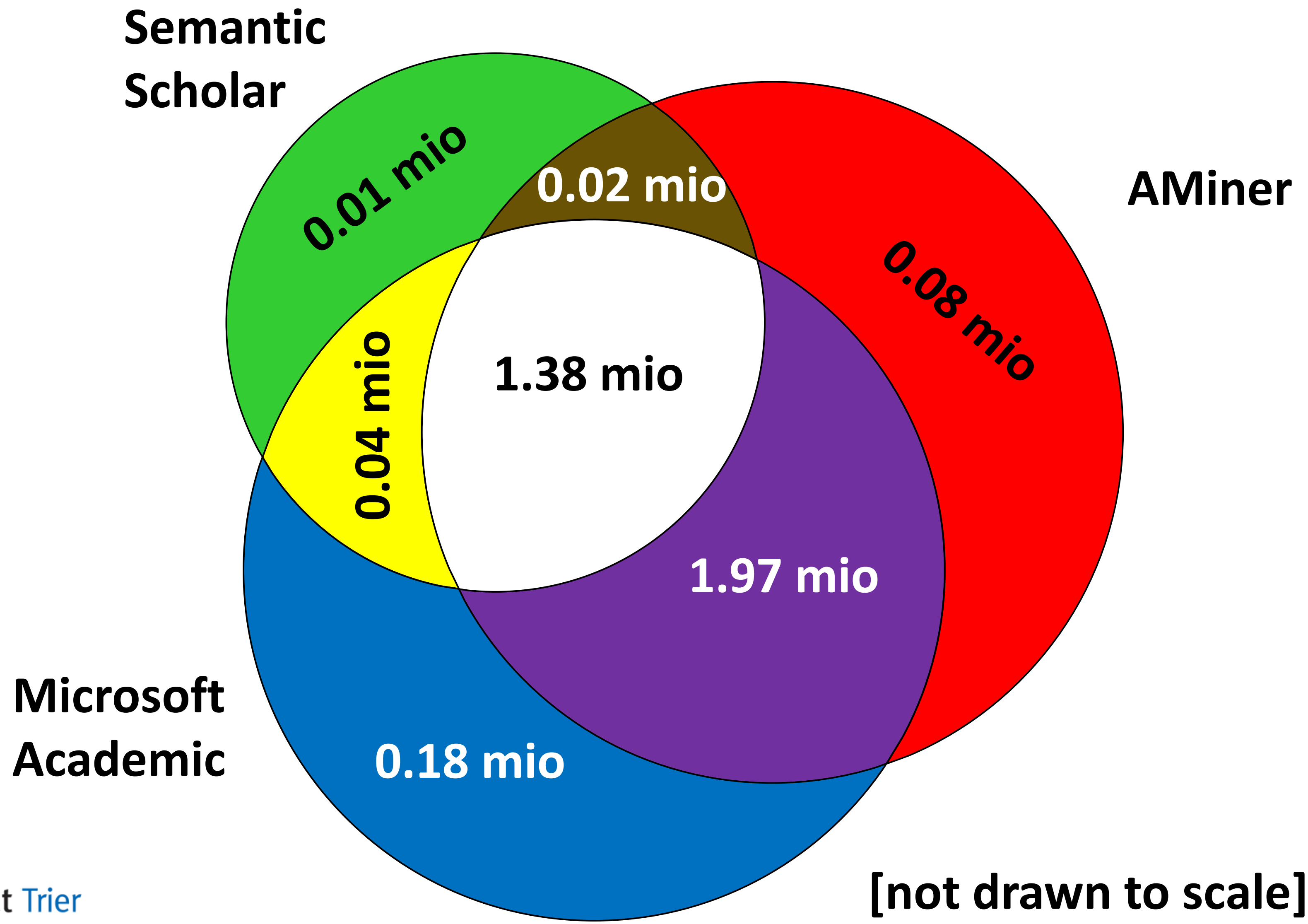
Post-Filter by author overlap, venue similarity, temporal proximity, …

dblp key
or
no match

**Mapping quality in general very good, no systematic evaluation yet**

# Coverage of dblp and Overlap

0.26 mio from dblp missing



[not drawn to scale]

# Overlap of dblp and CrossRef

DOI-based match in February 2018

- 4 million publications in dblp
- 3.2 million with DOI
- **3.1 million found in CrossRef**
- **600,000 with citations** (~15%)
  - 16 million citation instances
  - 4 million mapped based on DOI
  - ~1 million mapped based on reference string

## Main Observation:

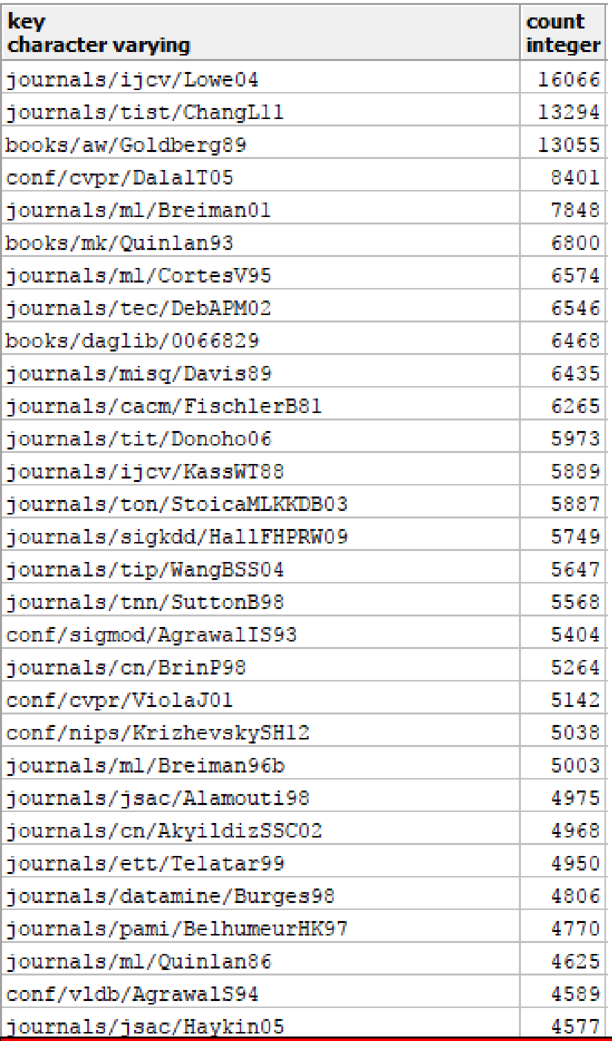
- All collections are **too incomplete** or **too static** to be useful for productive use.
- **Initiative for Open Citations** has effect, but still limited for computer science

# Outline

- Meta Data Harvesting
- Author Disambiguation
- Existing Metadata Collections
- **Citations**

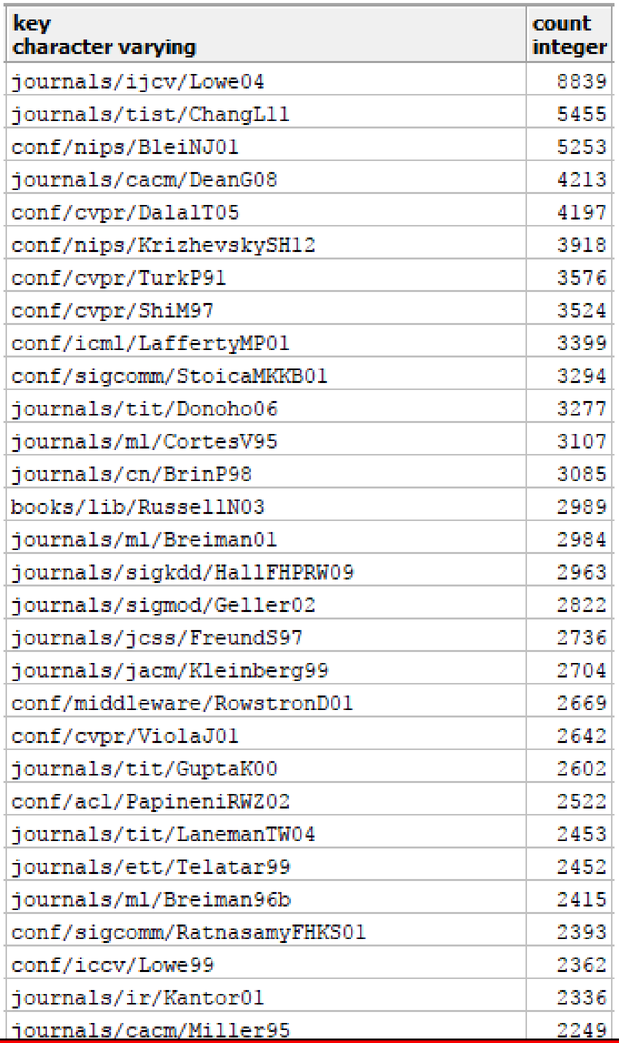
# Bibliometrics: most frequently cited pubs

## Aminer*



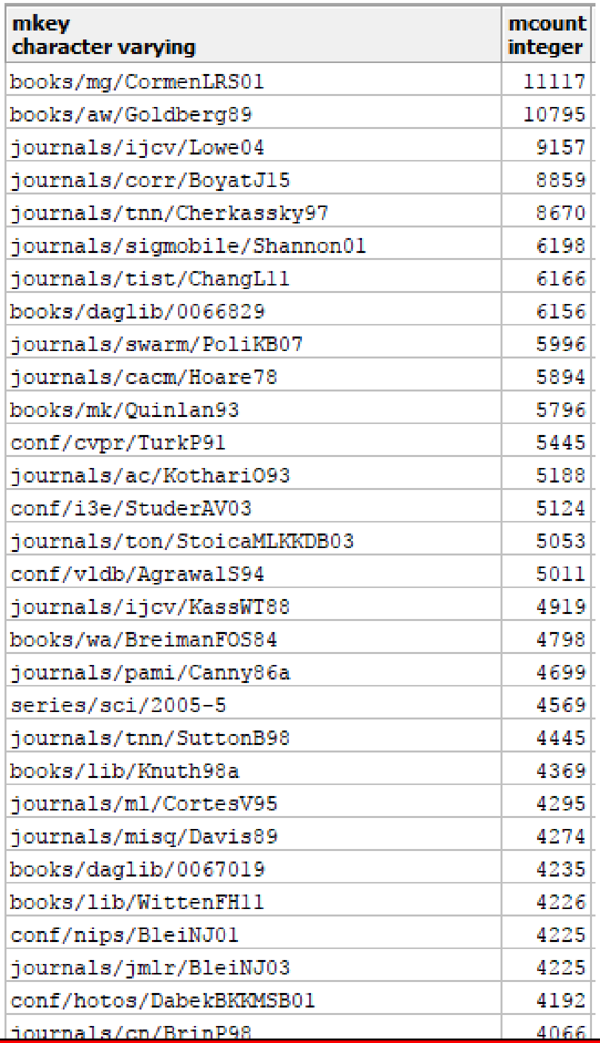
| key | count |
|---|---|
| character varying | integer |
| journals/ijcv/Lowe04 | 16066 |
| journals/tist/ChangL11 | 13294 |
| books/aw/Goldberg89 | 13055 |
| conf/cvpr/DalalT05 | 8401 |
| journals/ml/Breiman01 | 7848 |
| books/mk/Quinlan93 | 6800 |
| journals/ml/CortesV95 | 6574 |
| journals/tec/DebAPM02 | 6546 |
| books/daglib/0066829 | 6468 |
| journals/misq/Davis89 | 6435 |
| journals/cacm/FischlerB81 | 6265 |
| journals/tit/Donoho06 | 5973 |
| journals/ijcv/KassWT88 | 5889 |
| journals/ton/StoicaMLKKDB03 | 5887 |
| journals/sigkdd/HallFHPRW09 | 5749 |
| journals/tip/WangBSS04 | 5647 |
| journals/tnn/SuttonB98 | 5568 |
| conf/sigmod/AgrawalIS93 | 5404 |
| journals/cn/BrinP98 | 5264 |
| conf/cvpr/ViolaJ01 | 5142 |
| conf/nips/KrizhevskySH12 | 5038 |
| journals/ml/Breiman96b | 5003 |
| journals/jsac/Alamouti98 | 4975 |
| journals/cn/AkyildizSSC02 | 4968 |
| journals/ett/Telatar99 | 4950 |
| journals/datamine/Burges98 | 4806 |
| journals/pami/BelhumeurHK97 | 4770 |
| journals/ml/Quinlan86 | 4625 |
| conf/vldb/AgrawalS94 | 4589 |
| journals/jsac/Haykin05 | 4577 |

## SemanticScholar



| key | count |
|---|---|
| character varying | integer |
| journals/ijcv/Lowe04 | 8839 |
| journals/tist/ChangL11 | 5455 |
| conf/nips/BleiNJ01 | 5253 |
| journals/cacm/DeanG08 | 4213 |
| conf/cvpr/DalalT05 | 4197 |
| conf/nips/KrizhevskySH12 | 3918 |
| conf/cvpr/TurkP91 | 3576 |
| conf/cvpr/ShiM97 | 3524 |
| conf/icml/LaffertyMP01 | 3399 |
| conf/sigcomm/StoicaMKKB01 | 3294 |
| journals/tit/Donoho06 | 3277 |
| journals/ml/CortesV95 | 3107 |
| journals/cn/BrinP98 | 3085 |
| books/lib/RussellN03 | 2989 |
| journals/ml/Breiman01 | 2984 |
| journals/sigkdd/HallFHPRW09 | 2963 |
| journals/sigmod/Geller02 | 2822 |
| journals/jcss/FreundS97 | 2736 |
| journals/jacm/Kleinberg99 | 2704 |
| conf/middleware/RowstronD01 | 2669 |
| conf/cvpr/ViolaJ01 | 2642 |
| journals/tit/GuptaK00 | 2602 |
| conf/acl/PapineniRWZ02 | 2522 |
| journals/tit/LanemanTW04 | 2453 |
| journals/ett/Telatar99 | 2452 |
| journals/ml/Breiman96b | 2415 |
| conf/sigcomm/RatnasamyFHKS01 | 2393 |
| conf/iccv/Lowe99 | 2362 |
| journals/ir/Kantor01 | 2336 |
| journals/cacm/Miller95 | 2249 |

## Microsoft Academic*



| mkey | mcount |
|---|---|
| character varying | integer |
| books/mg/CormenLRS01 | 11117 |
| books/aw/Goldberg89 | 10795 |
| journals/ijcv/Lowe04 | 9157 |
| journals/corr/BoyatJ15 | 8859 |
| journals/tnn/Cherkassky97 | 8670 |
| journals/sigmobile/Shannon01 | 6198 |
| journals/tist/ChangL11 | 6166 |
| books/daglib/0066829 | 6156 |
| journals/swarm/PoliKB07 | 5996 |
| journals/cacm/Hoare78 | 5894 |
| books/mk/Quinlan93 | 5796 |
| conf/cvpr/TurkP91 | 5445 |
| journals/ac/Kothari093 | 5188 |
| conf/i3e/StuderAV03 | 5124 |
| journals/ton/StoicaMLKKDB03 | 5053 |
| conf/vldb/AgrawalS94 | 5011 |
| journals/ijcv/KassWT88 | 4919 |
| books/wa/BreimanFOS84 | 4798 |
| journals/pami/Canny86a | 4699 |
| series/sci/2005-5 | 4569 |
| journals/tnn/SuttonB98 | 4445 |
| books/lib/Knuth98a | 4369 |
| journals/ml/CortesV95 | 4295 |
| journals/misq/Davis89 | 4274 |
| books/daglib/0067019 | 4235 |
| books/lib/WittenFH11 | 4226 |
| conf/nips/BleiNJ01 | 4225 |
| journals/jmlr/BleiNJ03 | 4225 |
| conf/hotos/DabekBKKMSB01 | 4192 |
| journals/cn/BrinP98 | 4066 |

**Significantly different ranking derived from different collections – which one should one use?**

# Bibliometrics: Most prominent authors

## Aminer*

| Autor | #paper | #cites | h-Index |
|---|---|---|---|
| Jiawei Han 0001 | 828 | 33892 | 90 |
| Andrew Zisserman | 431 | 34226 | 86 |
| Anil K. Jain | 629 | 39336 | 86 |
| Scott Shenker | 301 | 33149 | 83 |
| Philip S. Yu | 1092 | 26027 | 79 |
| Hector Garcia-Molina | 455 | 20999 | 79 |
| Christos Faloutsos | 577 | 25758 | 76 |
| Sebastian Thrun | 305 | 21123 | 73 |
| Jitendra Malik | 248 | 29638 | 72 |
| Don Towsley | 626 | 18699 | 71 |
| Ion Stoica | 252 | 26341 | 71 |
| Andrew Y. Ng | 207 | 22617 | 71 |
| Thomas S. Huang | 946 | 25285 | 71 |
| Luc J. Van Gool | 686 | 29314 | 71 |
| Michael I. Jordan | 415 | 24734 | 70 |
| David E. Culler | 237 | 25451 | 69 |
| Georgios B. Giannakis | 745 | 18145 | 69 |
| Cordelia Schmid | 229 | 30525 | 68 |
| Francisco Herrera | 562 | 17041 | 68 |
| HongJiang Zhang | 396 | 18349 | 67 |

## Semantic Scholar*

| Autor | #paper | #cites | h-Index |
|---|---|---|---|
| Scott Shenker | 301 | 16287 | 65 |
| Andrew Y. Ng | 207 | 16971 | 57 |
| Hector Garcia-Molina | 455 | 11225 | 57 |
| Ion Stoica | 252 | 12819 | 55 |
| Jiawei Han 0001 | 828 | 15198 | 55 |
| Hari Balakrishnan | 196 | 15397 | 55 |
| Michael I. Jordan | 415 | 18203 | 54 |
| Yoshua Bengio | 446 | 14935 | 52 |
| Christos Faloutsos | 577 | 10405 | 51 |
| Deborah Estrin | 243 | 11958 | 51 |
| Sebastian Thrun | 305 | 8001 | 50 |
| Christopher D. Manning | 254 | 12180 | 50 |
| Andrew McCallum | 207 | 10308 | 49 |
| Daphne Koller | 257 | 7274 | 49 |
| Anil K. Jain | 629 | 10733 | 49 |
| Thomas A. Henzinger | 449 | 9391 | 49 |
| Alon Y. Halevy | 253 | 8021 | 49 |
| Jennifer Widom | 216 | 8254 | 48 |
| Jon M. Kleinberg | 269 | 10906 | 47 |
| Joseph M. Hellerstein | 214 | 7684 | 47 |

**Significantly different ranking derived from different collections – which one should one use?**

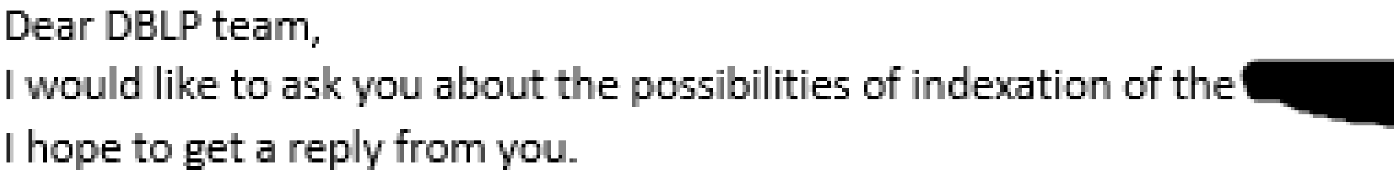
# Scientific Challenge:

**Make bibliometric measures aware of incompleteness and possible errors**

**Provide confidence intervals for bibliometric measures**

# Possible uses of citations in dblp

- Estimate **importance of conferences** (to decide if and when a conference should be added)



Dear DBLP team,

I would like to ask you about the possibilities of indexation of the [redacted]

I hope to get a reply from you.

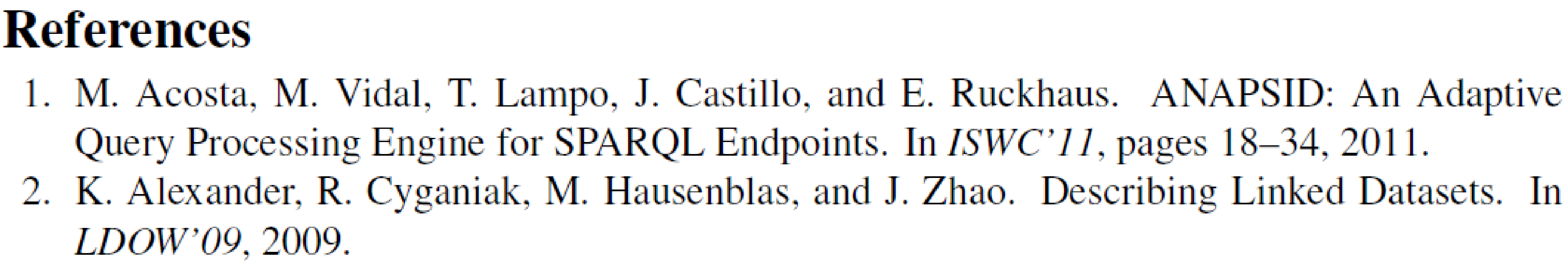
- Identify publication venues where **coverage in dblp is incomplete** (and missing part is important)
- Identify important **new publication venues**

# DIY-Extraction from PDFs

- **ScienceParse** by Allen Institute for AI
- Reads (OCR‘ed) PDF as input
- Yields
  - Abstract
  - Authors with Emails
  - Full text with (some) structure
  - Citations with (some) structure

**https://github.com/allenai/science-parse**

# Citations

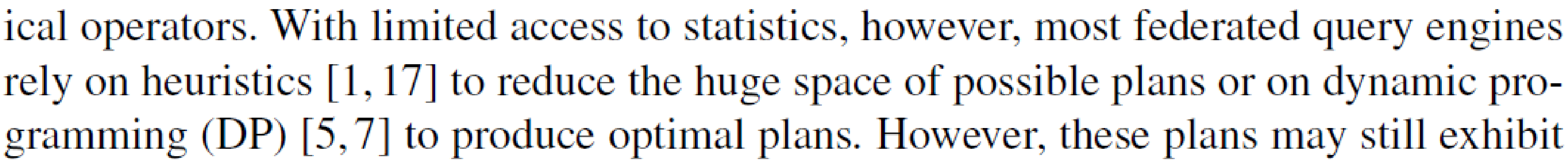


## References

1. M. Acosta, M. Vidal, T. Lampo, J. Castillo, and E. Ruckhaus. ANAPSID: An Adaptive Query Processing Engine for SPARQL Endpoints. In *ISWC'11*, pages 18–34, 2011.
2. K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. Describing Linked Datasets. In *LDOW'09*, 2009.

```
"references" : [ {
    "title" : "ANAPSID: An Adaptive Query Processing Engine for SPARQL Endpoints",
    "author" : [ "M. Acosta", "M. Vidal", "T. Lampo", "J. Castillo", "E. Ruckhaus"],
    "venue" : "In ISWC'11,",
    "citeRegEx" : "1",
    "shortCiteRegEx" : "1",
    "year" : 2011
  }, {
    "title" : "Describing Linked Datasets",
    "author" : [ "K. Alexander", "R. Cyganiak", "M. Hausenblas", "J. Zhao" ],
    "venue" : "In LDOW'09,",
    "citeRegEx" : "2",
    "shortCiteRegEx" : "2",
    "year" : 2009
  }, …
```

# Citation Contexts



ical operators. With limited access to statistics, however, most federated query engines rely on heuristics [1, 17] to reduce the huge space of possible plans or on dynamic programming (DP) [5, 7] to produce optimal plans. However, these plans may still exhibit

```
"referenceMentions" : [ {
      "referenceID" : 0,
      "context" : "Federated SPARQL query engines [1, 4, 7, 14, 17] answer SPARQL queries over a
federation of SPARQL endpoints.",
      "startOffset" : 31,
      "endOffset" : 48
    }, {
      "referenceID" : 0,
      "context" : "With limited access to statistics, however, most federated query engines rely
on heuristics [1, 17] to reduce the huge space of possible plans or on dynamic programming (DP)
[5, 7] to produce optimal plans.",
      "startOffset" : 92,
      "endOffset" : 99
    }, …
```

# Evaluating Mapping Quality for Citations

96 papers from PVLDB Volume 10

- 3084 manually annotated citations
- 2700 with well-defined match in dblp

**Results:** (with best parameter setting, no systematic eval)

- Recall: ~80%
- Precision: ~97.5%
- Accuracy of match/nonmatch decisions: ~81%

**A lot worse on old, OCR‘ed publications until ~2000 (finding citation & segmentation fails, OCR errors, …)**

# Experiment on CoRR Jan-Jun 2017

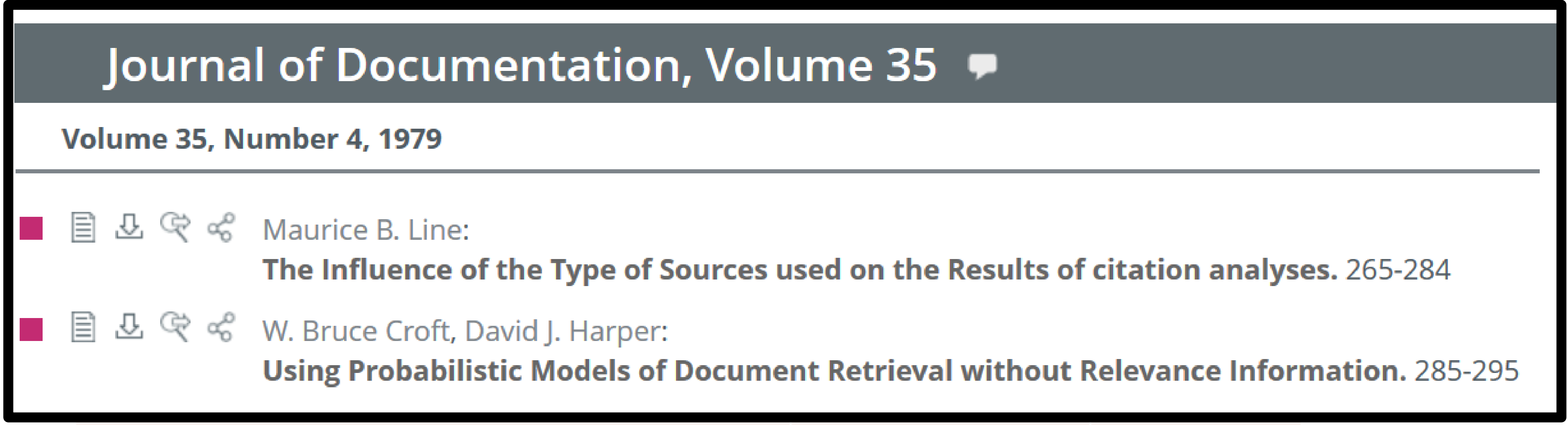
## Most frequently extracted venues (after some normalization)

| venue | matched | not matched | overall | missing | found |
|---|---:|---:|---:|---:|---:|
| cvpr | 5120 | 47 | 5167 | 0,91% | 99,09% |
| advances in neural information processing systems | 4205 | 66 | 4271 | 1,55% | 98,45% |
| nips | 2795 | 60 | 2855 | 2,10% | 97,90% |
| ieee conference on computer vision and pattern recognition | 2806 | 43 | 2849 | 1,51% | 98,49% |
| corr | 2327 | 45 | 2372 | 1,90% | 98,10% |
| ieee transactions on information theory | 2004 | 71 | 2075 | 3,42% | 96,58% |
| ieee trans. inf. theory | 2005 | 61 | 2066 | 2,95% | 97,05% |
| iccv | 1807 | 27 | 1834 | 1,47% | 98,53% |
| eccv | 1809 | 14 | 1823 | 0,77% | 99,23% |
| journal of machine learning research | 1519 | 281 | 1800 | 15,61% | 84,39% |
| icml | 1714 | 70 | 1784 | 3,92% | 96,08% |
| phd thesis | 577 | 1160 | 1737 | 66,78% | 33,22% |
| ieee transactions on pattern analysis and machine intelligence | 1553 | 69 | 1622 | 4,25% | 95,75% |
|  | 464 | 1146 | 1610 | 71,18% | 28,82% |
| international conference on machine learning | 1486 | 46 | 1532 | 3,00% | 97,00% |
| ieee trans. wireless commun | 1328 | 62 | 1390 | 4,46% | 95,54% |
| technical report | 235 | 1035 | 1270 | 81,50% | 18,50% |
| ieee | 868 | 350 | 1218 | 28,74% | 71,26% |
| ieee trans. signal process | 1049 | 45 | 1094 | 4,11% | 95,89% |
| neural computation | 1046 | 40 | 1086 | 3,68% | 96,32% |
| ieee transactions on signal processing | 949 | 73 | 1022 | 7,14% | 92,86% |
| ieee transactions on automatic control | 806 | 197 | 1003 | 19,64% | 80,36% |

# Experiment on CoRR Jan-Jun 2017

## Venues with significant holes in dblp

| Venue | found | not found |
|---|---:|---:|
| phd thesis | 577 | 1160 |
|  | 464 | 1146 |
| technical report | 235 | 1035 |
| science | [illegible] | 578 |
| springer | 195 | 320 |
| journal of machine learning research | 1519 | 281 |
| ieee transactions on power systems | 34 | 280 |
| physical review letters | 54 | 268 |
| crc press | 36 | 203 |
| ieee transactions on automatic control | 806 | 197 |
| master’s thesis | 17 | 185 |



## Journal of Documentation, Volume 35

Volume 35, Number 4, 1979

Maurice B. Line:
**The Influence of the Type of Sources used on the Results of citation analyses.** 265-284

W. Bruce Croft, David J. Harper:
**Using Probabilistic Models of Document Retrieval without Relevance Information.** 285-295

# Experiment on CoRR Jan-Jun 2017

## Venues that could not be matched to dblp

| venue | Not found |
|---|---:|
| the annals of mathematical statistics | 168 |
| psychological review | 152 |
| journal of the royal statistical society. series b | 139 |
| journal of personality and social psychology | 96 |
| journal of statistical software | 85 |
| american journal of sociology | 69 |
| behavior research methods | 68 |
| econometrica: journal of the econometric society | 64 |
| biglearn | [illegible] |
| wiley online library | 53 |
| naval research logistics quarterly | 49 |
| cognitive psychology | 46 |
| the journal of physiology | 45 |
| annual review of sociology | 45 |
| journal of marketing research | 44 |
| monthly weather review | 44 |
| mathematische annalen | 43 |
| problemy peredachi informatsii | 42 |
| biometrics | 40 |

⇦ Missing NIPS workshop (no longer available)

Math

Sociology

Psychology

Other Sciences

# Conclusion

- **Open meta data** is becoming more important and more available
- **Quality and scope** of available meta data is still unclear
- **Bibliometric measures** must take this uncertainty into account

# Future Work for dblp

- Integrate with more **data providers** (currently ORCID and WikiData)
- Connect to **bibliographic data providers** from other domains
- Develop **model for conference series** and events
- Include references to **published data** (e.g., DataCite)

# Future Work for Research

- Collect more **extensive metadata for conferences**
  - Organizers
  - Members of the program committee
  - Reviewers
  - Keynote speakers
  - …
- Exploit this information for better **estimation of the reputation** of scientists (and of conferences)