Bibliographic Information Sources for Computer Science with a Focus on Citations

Ralf Schenkel

dblp
computer science bibliography

Universität Trier

Databases & Information Systems Group

- Prof. Dr.-Ing. Ralf Schenkel

  2003-2013 Max-Planck-Institut für Informatik, Saarbrücken
  2013-2016 Universität Passau
  2016- Universität Trier

- Dr. Michael Ley

- Christin Kreutz

- Tobias Zeimetz

- Christopher Michels DFG
- Lorik Dumani DFG

Dblp core team

- Dr. Michael Ley 
- Dr. Marcel R. Ackermann 
- Oliver Hoffmann 
- Dr. Florian Reitz 
- Dr. Michael Wagner 
- Stefanie von Keutz 

maintained by SCHLOSS DAGSTUHL at Universität Trier

Dblp Overview

ACM/IEEE Joint Conference on Digital Libraries (JCDL)

> Home > Conferences and Workshops

Trier 1

www.jcdl.org

ACM DL - IEEE ADL

18. JCDL 2018: Fort Worth, TX, USA

Jiangping Chen, Marcos André Gonçalves, Jeff M. Allen, Edward A. Fox, Min-Yen Kan, Vivien Petras:
Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, JCDL 2018, Fort Worth, TX, USA, June 03-07, 2018. ACM 2018 [contents]

17. JCDL 2017: Toronto, ON, Canada

2017 ACM/IEEE Joint Conference on Digital Libraries, JCDL 2017, Toronto, ON, Canada, June 19-23, 2017. IEEE Computer Society 2017, ISBN 978-1-5386-3861-3 [contents]

Proceedings of the 6th International Workshop on Mining Scientific Publications, WOSP@JCDL 2017, Toronto, ON, Canada, June 19, 2017. ACM 2017 [contents]

16. JCDL 2016: Newark, NJ, USA

Nabil R. Adam, Lillian (Boots) Cassel, Yelena Yesha, Richard Furuta, Michele C. Weigle:
Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, JCDL 2016, Newark, NJ, USA, June 19 - 23, 2016. ACM 2016, ISBN 978-1-4503-4229-2 [contents]

Paul D. Clough, Paula Goodale, Maristella Agosti, Séamus Lawless:
Proceedings of the First International Workshop on Accessing Cultural Heritage at Scale co-located with Joint Conference on Digital Libraries 2016 (JCDL 2016), Newark, USA, June 22, 2016. CEUR Workshop Proceedings 1611, CEUR-WS.org 2016 [contents]

Guillaume Cabanac, Muthu Kumar Chandrasekaran, Ingo Frommholz, Kokil Jaidka, Min-Yen Kan, Philipp Mayr, Dietmar Wolfram:
Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL) co-located with the Joint Conference on Digital Libraries 2016 (JCDL 2016), Newark, NJ, USA, June 23, 2016. CEUR Workshop Proceedings 1610, CEUR-WS.org 2016 [contents]

Dblp Overview

18th JCDL 2018: Fort Worth, TX, USA

Jiangping Chen, Marcos André Gonçalves, Jeff M. Allen, Edward A. Fox, Min-Yen Kan, Vivien Petras:
Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, JCDL 2018, Fort Worth, TX, USA, June 03-07, 2018. ACM 2018

Keynote Talks

Trevor Owens:
We Have Interesting Problems: Some Applied Grand Challenges from Digital Libraries, Archives and Museums. 1

Niall Gaffney:
Improving Research Outcomes Leveraging Digital Libraries, Advanced Computing, and Data. 3

Carly Strasser:
Open Source Tech for Scholarly Communication: Why It Matters. 5

Session 1A: Use

Lucy McKenna, Christophe Debruyne, Declan O'Sullivan:
Understanding the Position of Information Professionals with regards to Linked Data: A Survey of Libraries, Archives and Museums. 7-16

Samuel Dodson, Luanne Freund, Rick Kopak:
The Role of Pre-Existing Highlights in Reader-Text Interactions and Outcomes. 17-20

Sampath Jayarathna, Sobiga Shanmugathasan:
Evaluating Saccade-Bounded Eye Movement Features for the User Modeling. 21-24

Shengli Deng, Jingjing Tong, Shaoxiong Fu:
Interaction on An Academic Social Networking Sites: A Study of ResearchGate Q&A on Library and Information Science. 25-28

Session 1B: Collection Building

Myriam C. Traub, Thaer Samar, Jacco van Ossenbruggen, Lynda Hardman:
Impact of Crowdsourcing OCR Improvements on Retrievability Bias. 29-36

Dblp

[+] [–] Ralf Schenkel

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

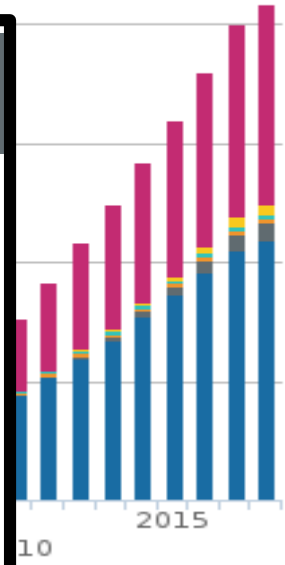
authority control:

> Home > Persons

[–] **Person information**

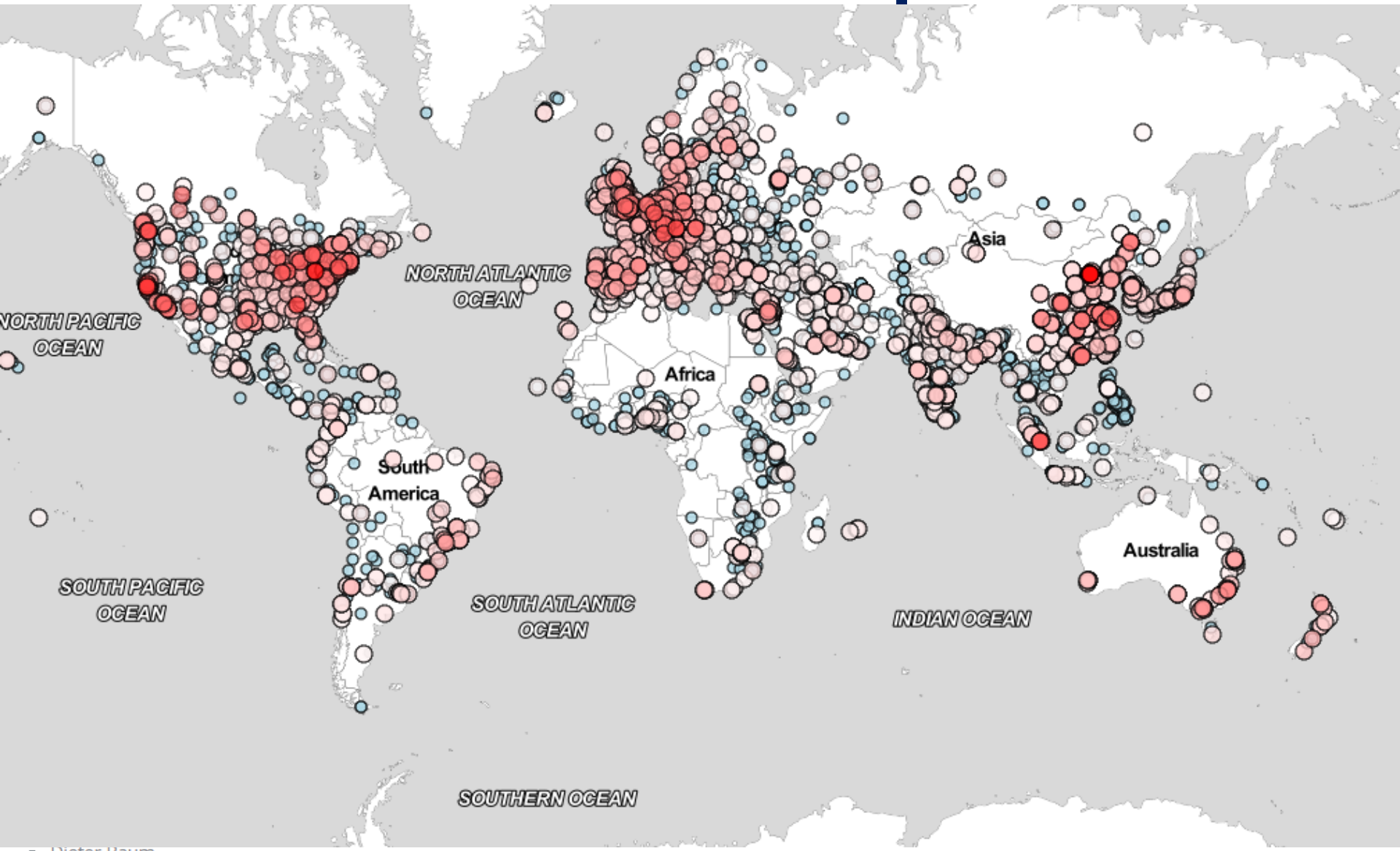
- *affiliation:* University of Trie
- *affiliation (former):* Universit
- *affiliation (former):* Saarland University, Saarbrücken, Germany

Recent activities:
*** affiliation information**
*** links to other authority providers, esp. ORCID and WikiData**



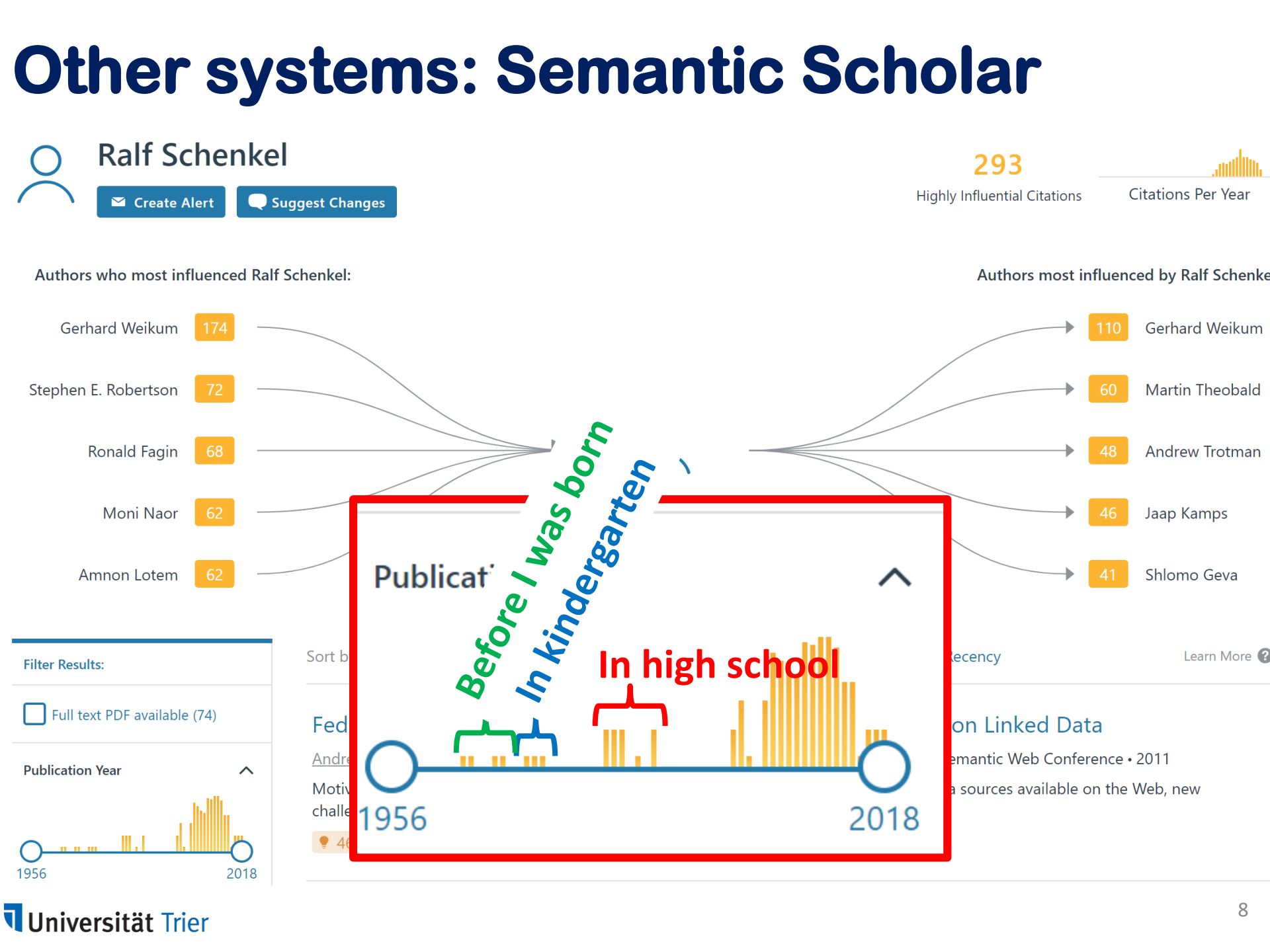
~4.2 million publications, ~2.1 million authors, ~400.000 new publications per year

Preview: Affiliations in dblp



- Dieter Baum
  University of Trier, Germany
- Fabian Beck 0001
  University of Trier, Computer Science Department, Germany
- Rainer Benda

Other systems: Semantic Scholar



Other Systems: MS Academic

Microsoft Academic

ralf schenkel

Ralf Schenkel

Saarland University

Fields of Study: Computer Science, Information retrieval, Data mining, Database, Query optimization, Ranking, Query expansion, XML, Indexation, Web search query, ...

Papers (189) Citations (4,124) Claim

Other authors with same name:

Ralf Schenkel

Top co-author: Arnab Kumar Dutta...

Top paper: A Distributed In-Memory SPARQL Quer...

Ralf Schenkel

Top co-author: Marcin Sydow...

Top paper: QBEES: Query-by-Example Entity Sear...

?

Ralf Schenkel

Top co-author: Jerome Galy...

Top paper: Nuclear reactions triggered by rela...

Ralf Schenkel

Top paper: Integrating and Exploiting Public M...

Other Systems: Google Scholar

Michael Ley

FOLGEN

University of Trier, DBLP
Bestätigte E-Mail-Adresse bei uni-trier.de
Digital Libraries Information Retrieval Database Systems

TITEL	ZITIERT VON	JAHR
Ontologies improve text document clustering A Hotho, S Staab, G Stumme Data Mining, 2003. ICDM 2003. Third IEEE International Conference on, 541-544	698	2003
The DBLP computer science bibliography: Evolution, research issues, perspectives M Ley International symposium on string processing and information retrieval, 1-10	305	2002
DBLP: some lessons learned M Ley Proceedings of the VLDB Endowment 2 (2), 1493-1500	285	2009
DBLP computer science bibliography M Ley University of Trier	99	2005
Maintaining an Online Bibliographical Database: The Problem of Data Quality. M Ley, P Reuther EGC, 5-10	76	2006

Other Systems

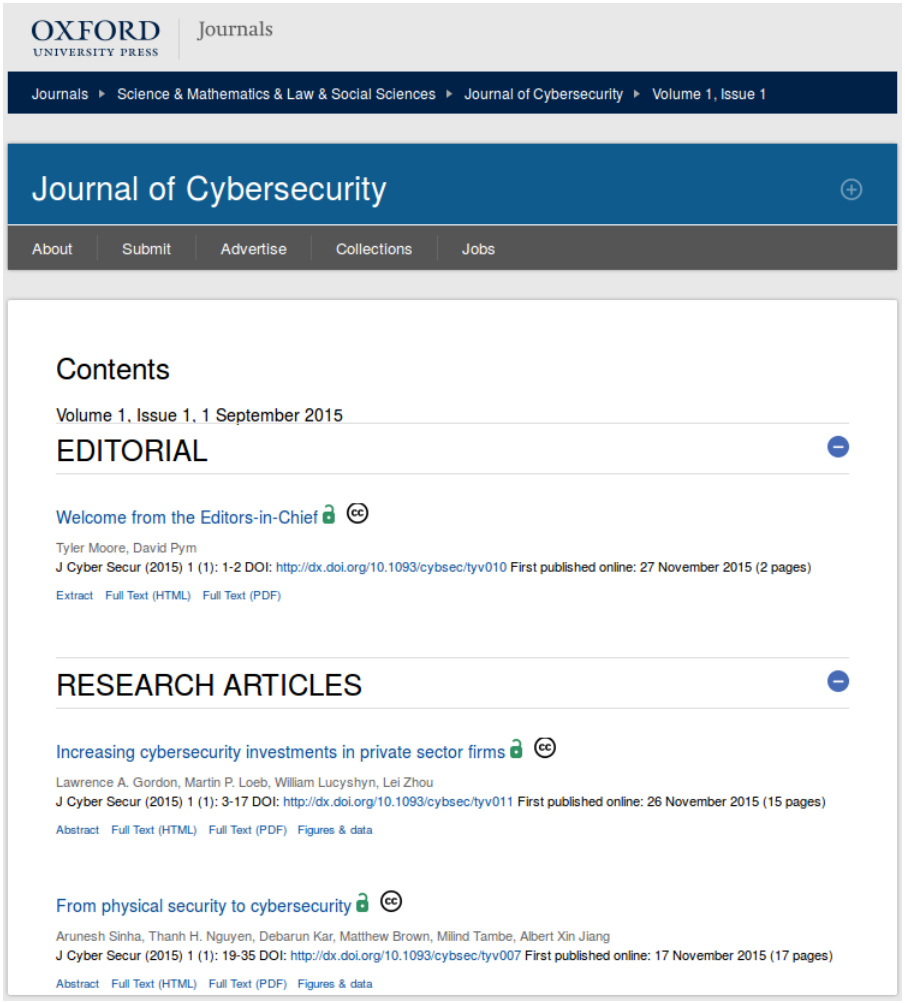
Common weakness:

Data quality issues due to automatic information collection

Advantage of dblp:

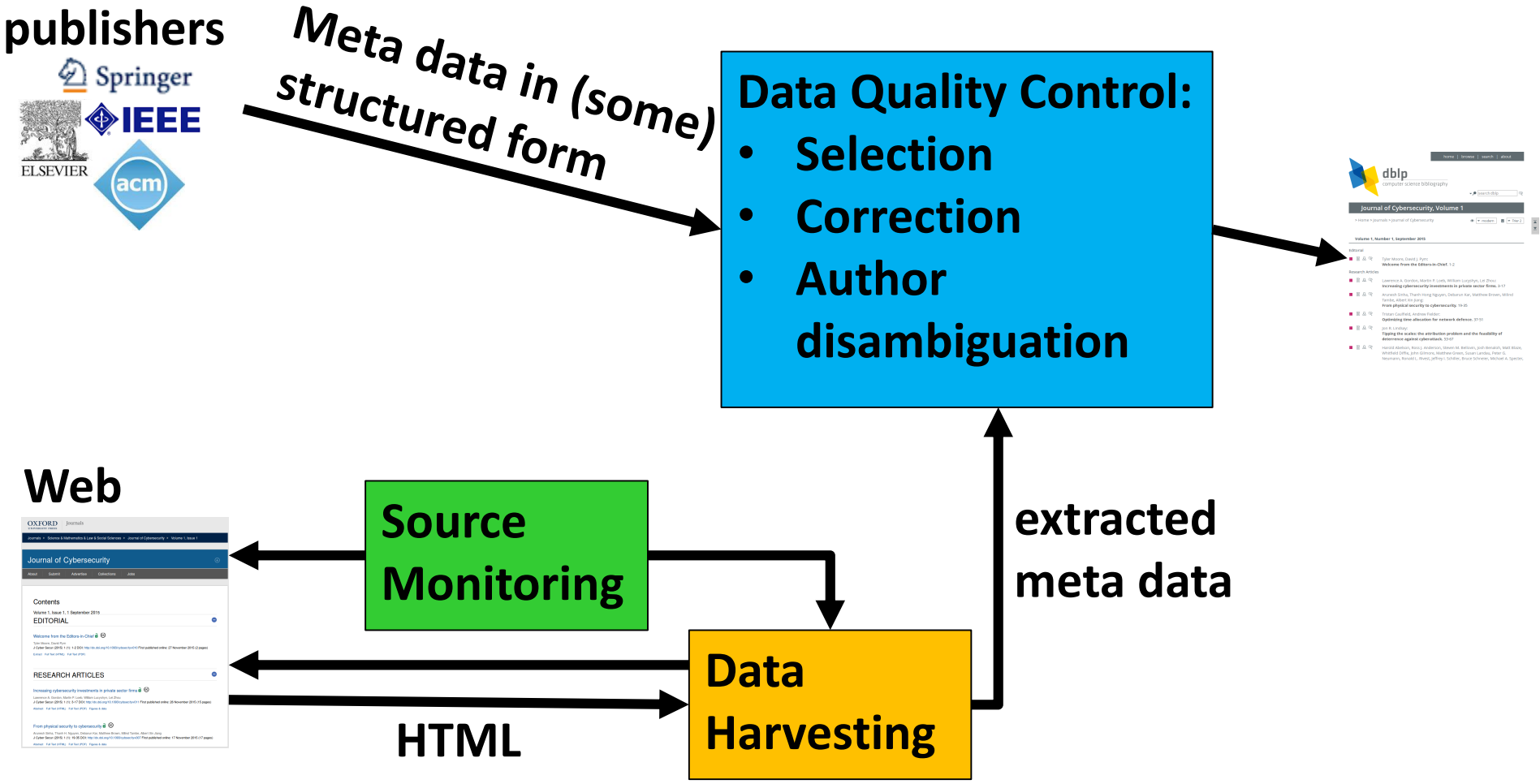
Manual data curation (with limits)

How is publication data added to dblp?



?

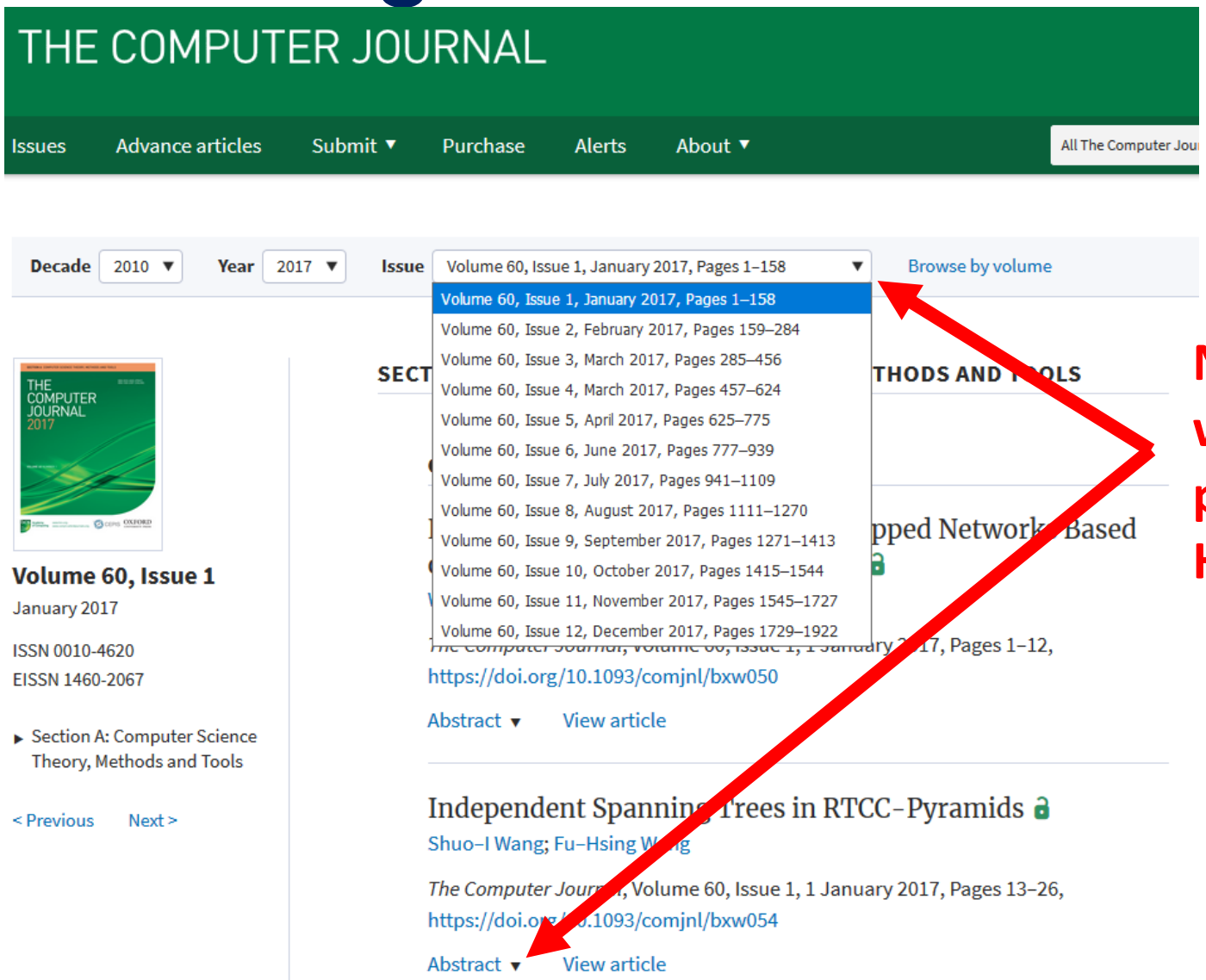
Dblp Data Ingestion Pipeline



Outline

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

Harvesting is much more difficult now



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

Successful harvesting needs to implement Javascript

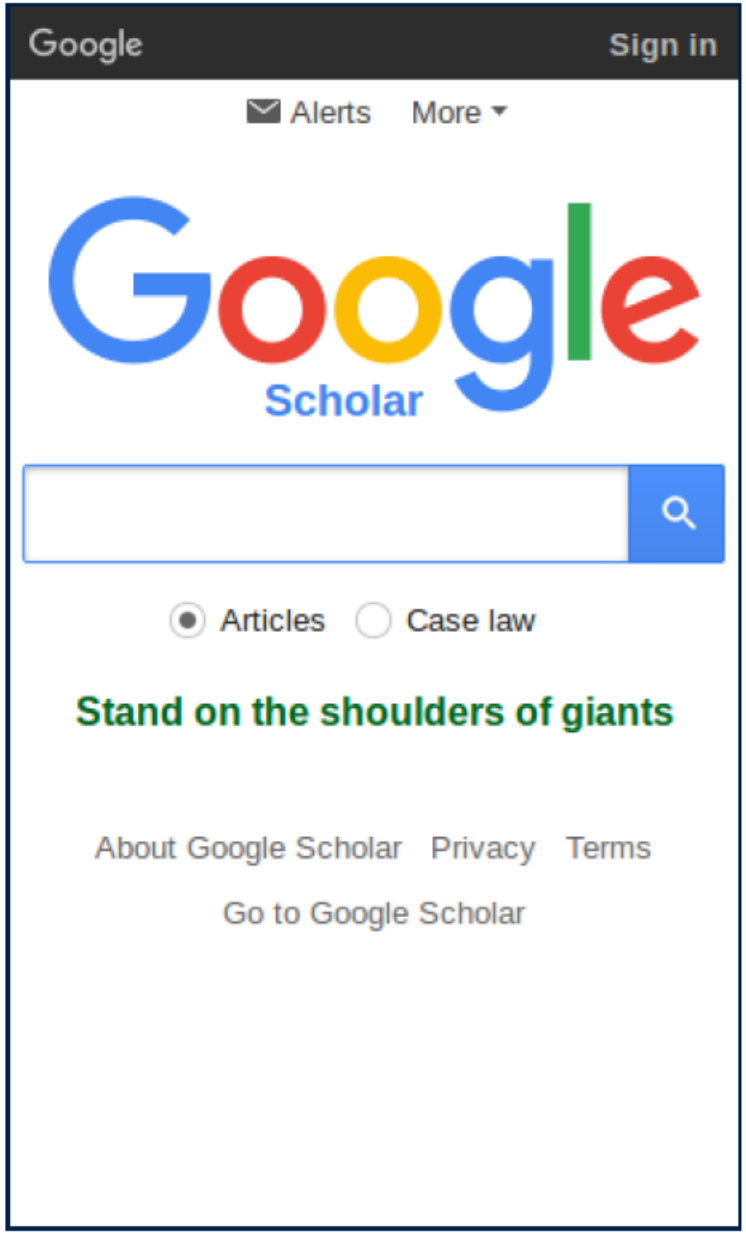
Monitoring & harvesting: OXPath

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

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

Example: Navigating Google Scholar

OXPath Expression

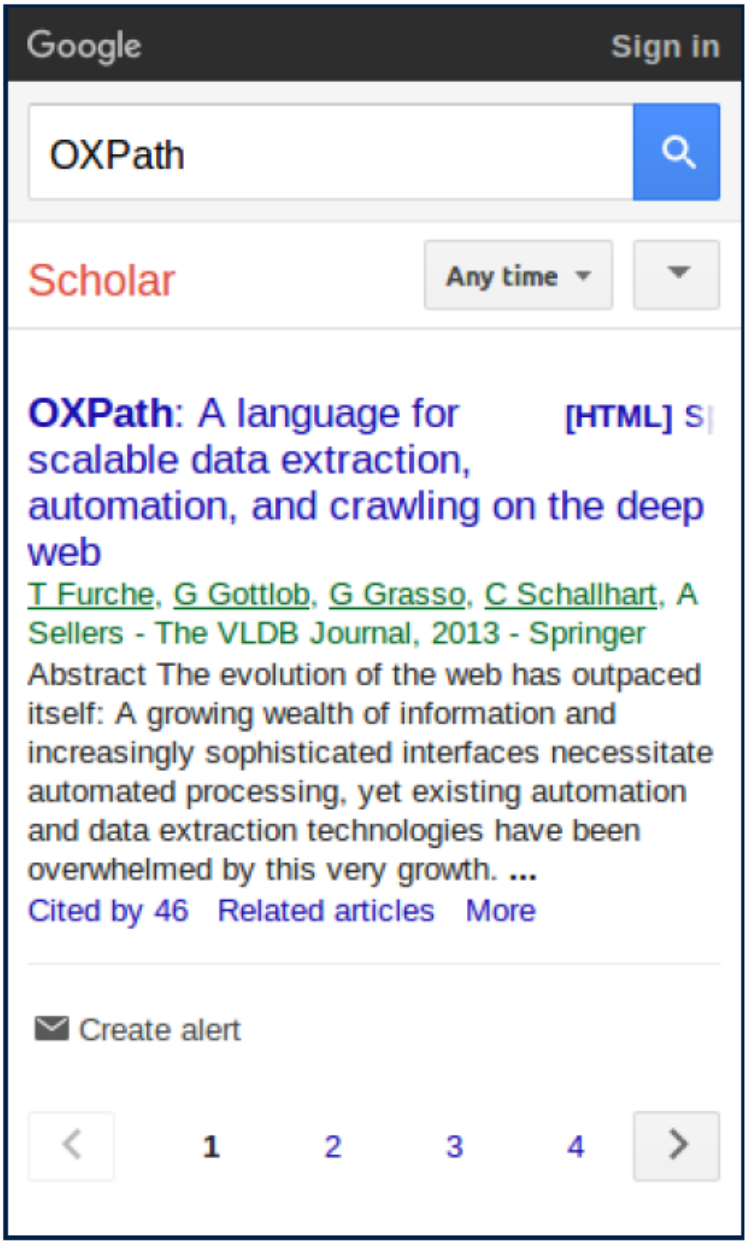
```
1 doc("https://scholar.google.com")
2   //*[@role="search"]//input[@type="text"]/{"OXPath"}
```

Example: Navigating Google Scholar

OXPath Expression

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

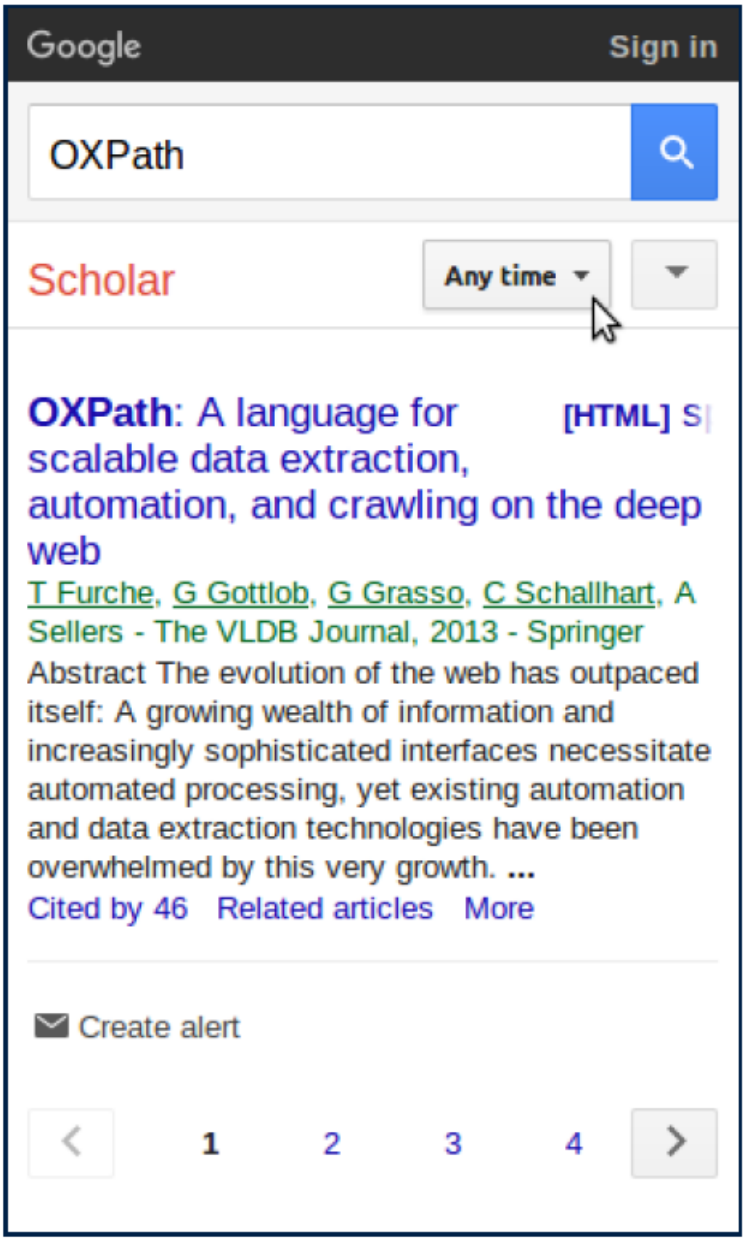
Example: Navigating Google Scholar



OXPath Expression

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

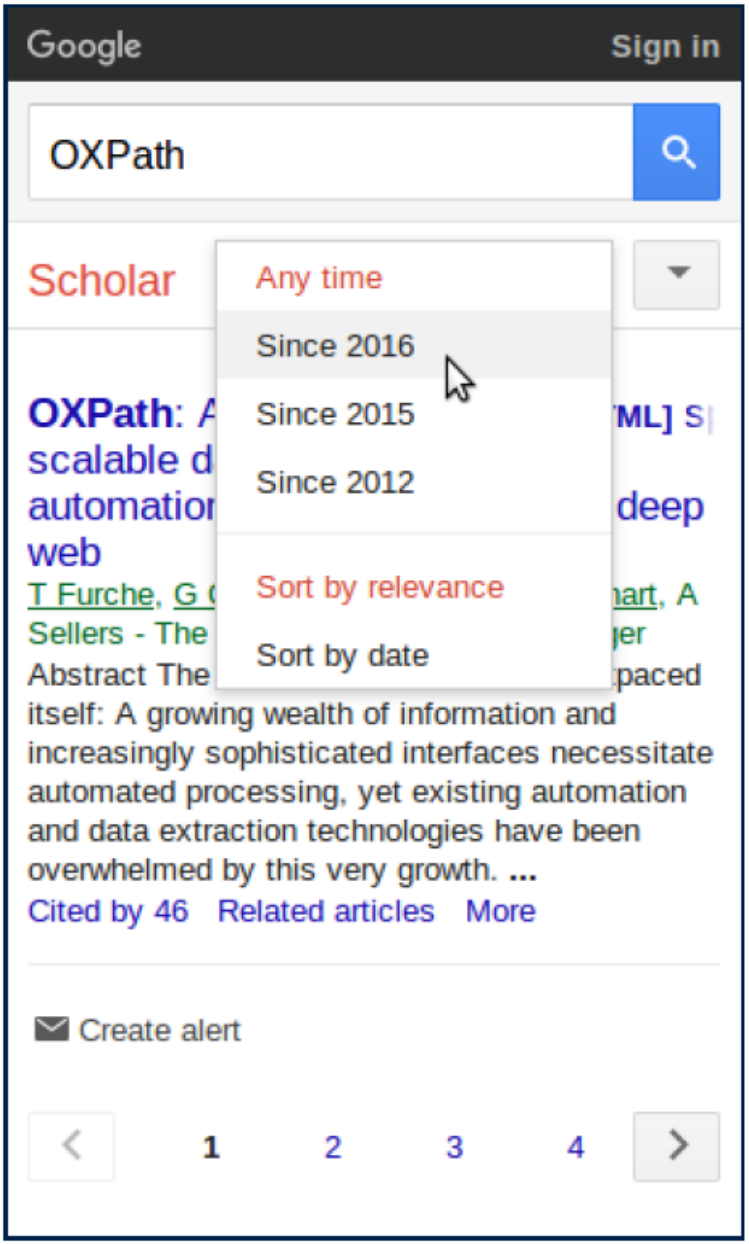
Example: Navigating Google Scholar



OXPath Expression

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

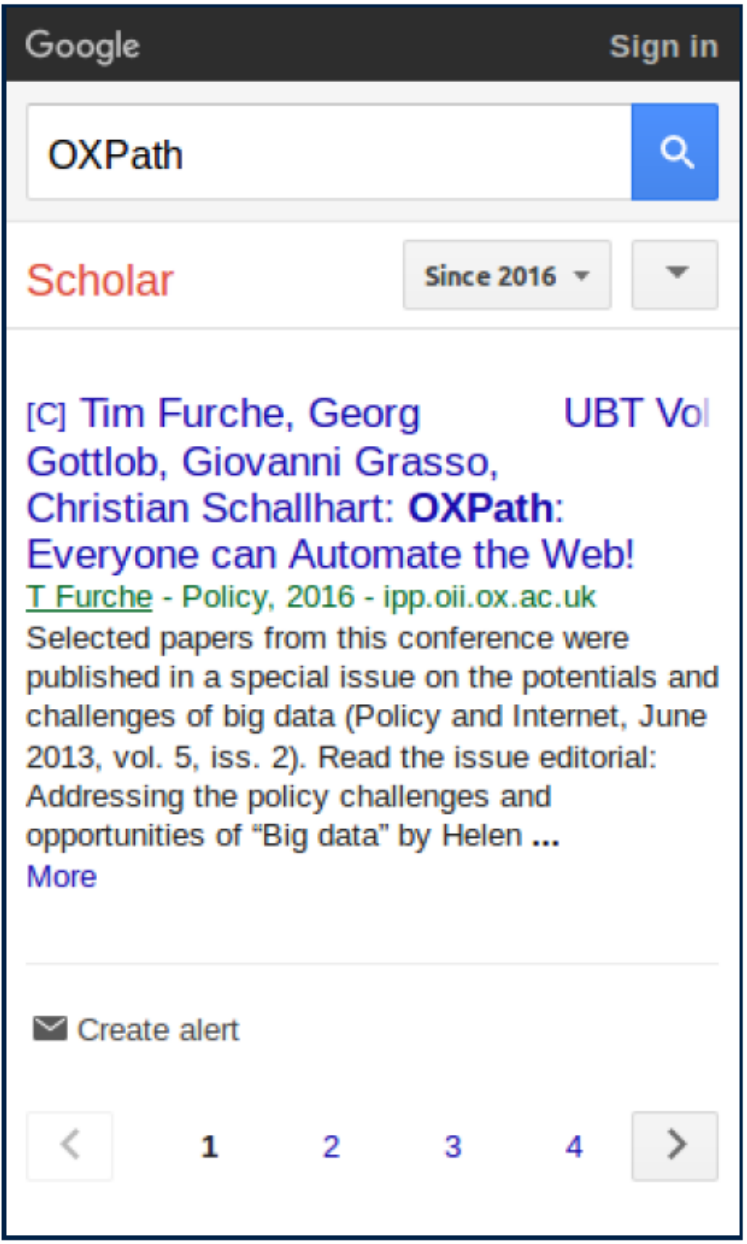
Example: Navigating Google Scholar



OXPath Expression

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

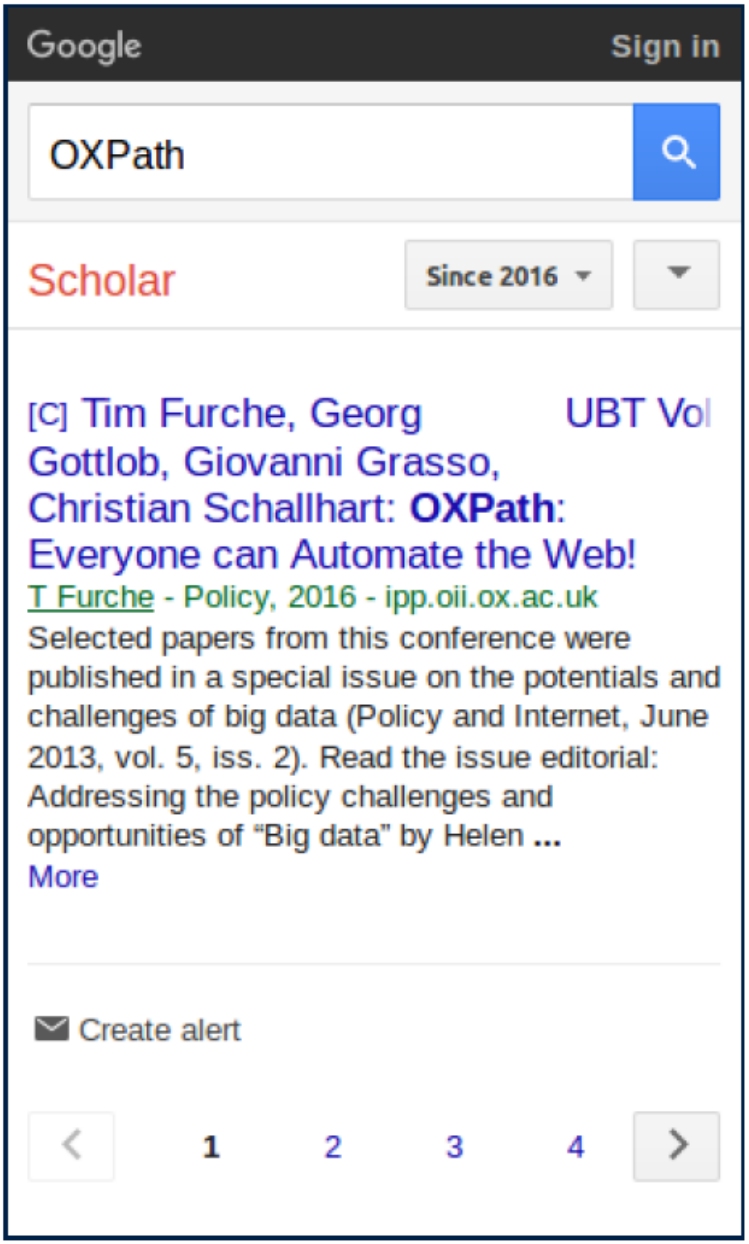
Example: Navigating Google Scholar



OXPath Expression

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

Example: Navigating Google Scholar



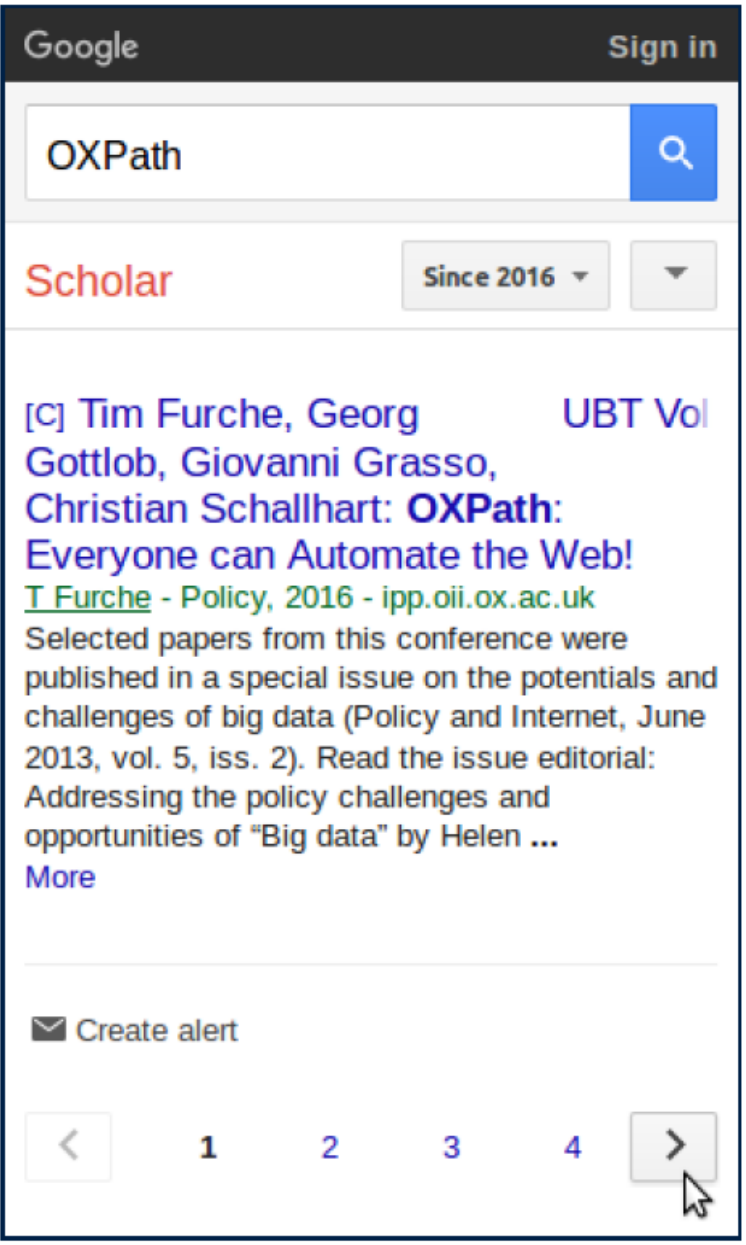
OXPath Expression

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

XML Output

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

Example: Navigating Google Scholar



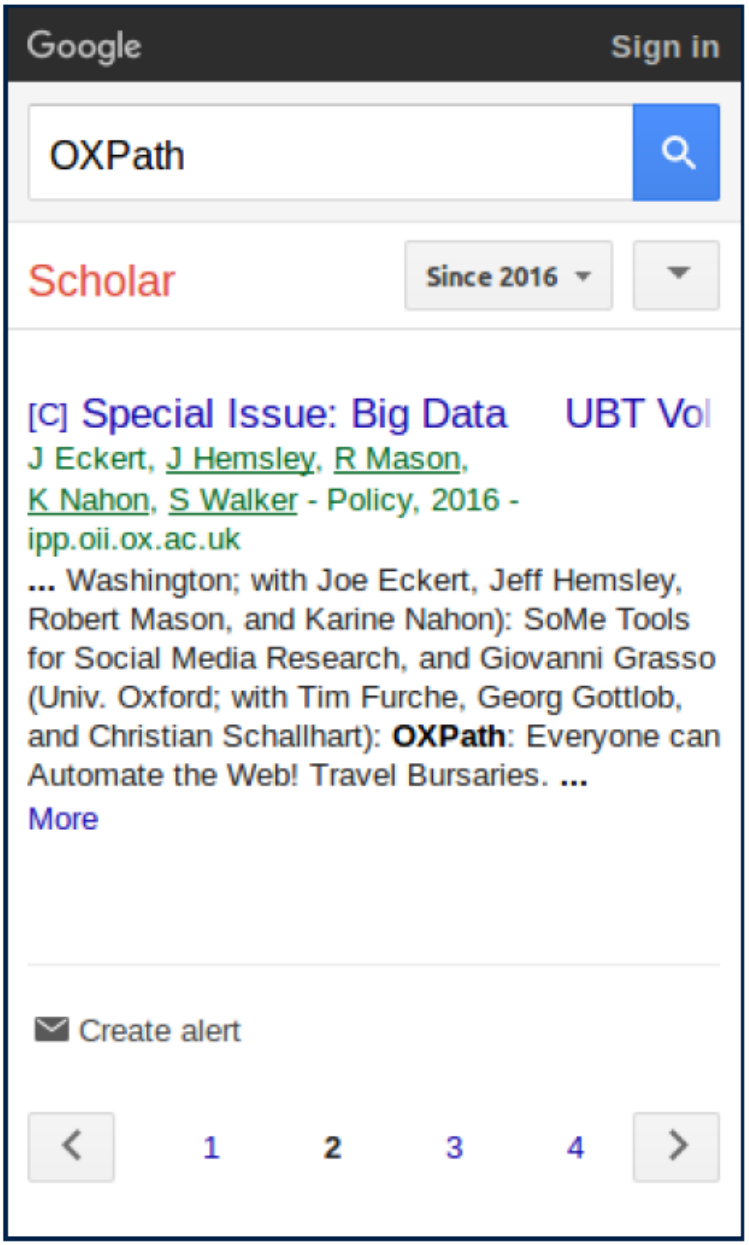
OXPath Expression

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

XML Output

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

Example: Navigating Google Scholar



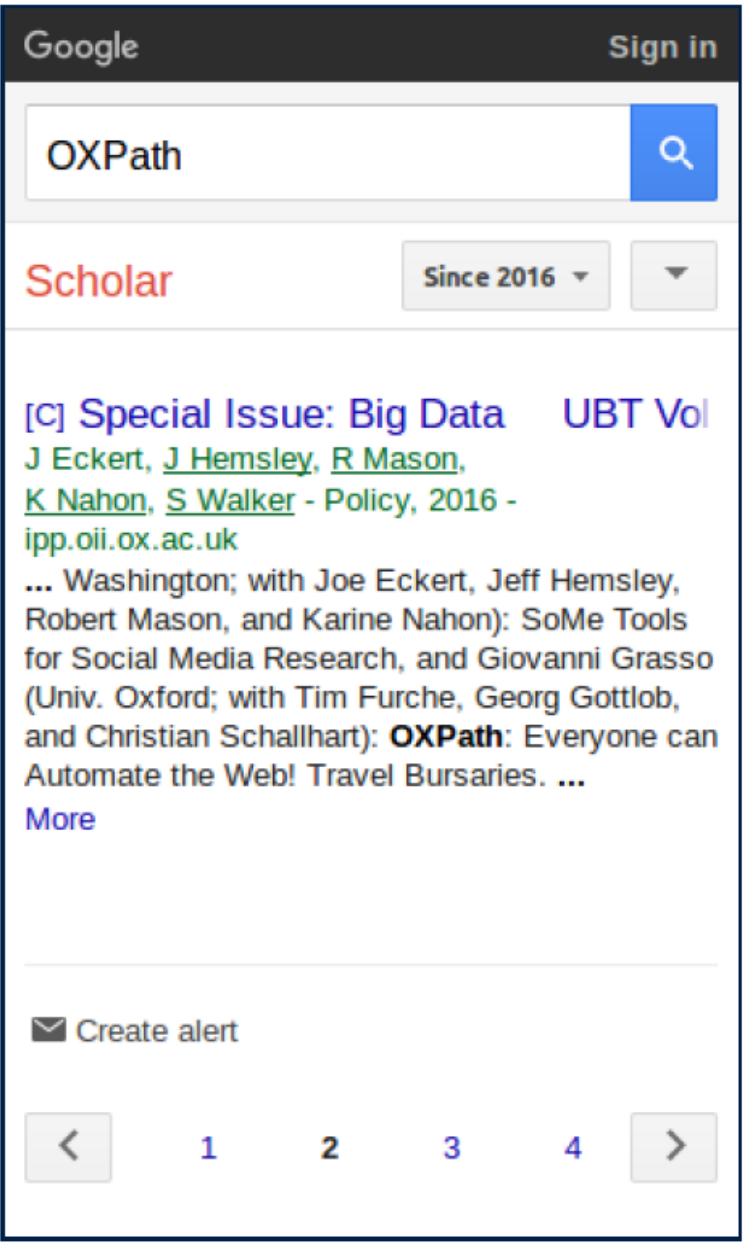
OXPath Expression

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

XML Output

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

Example: Navigating Google Scholar



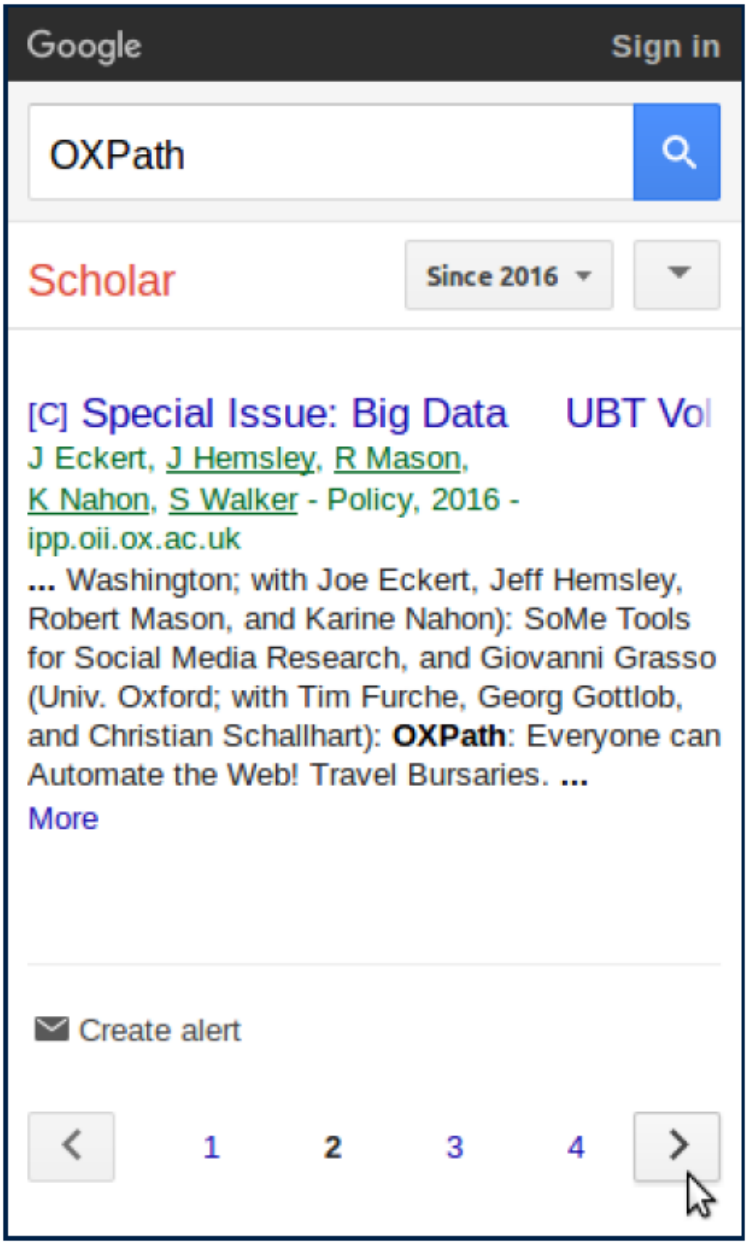
OXPath Expression

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

XML Output

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

Example: Navigating Google Scholar



OXPath Expression

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

XML Output

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <results>
3   <title>Tim Furche, Georg Gottlob, [...]</title>
4   <title>Special Issue: Big Data [...]</title>
5   <!--[...]-->
6 </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**

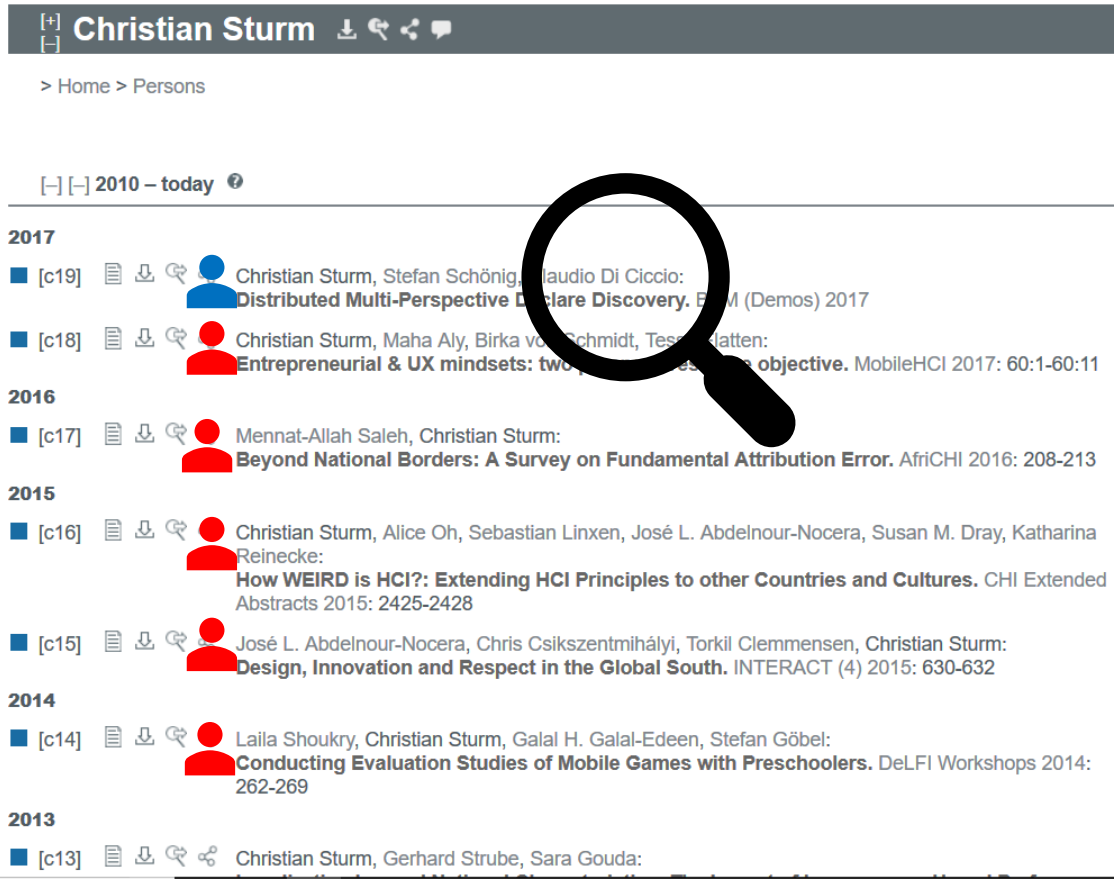
Now in productive use at dblp

Outline

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

Author Disambiguation: Homonyms

Multiple persons with the same name in the same profile



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

Affiliations would be useful, but usually not available

[+] [–] 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

[+] [–] Christian Sturm 0003
Christian Andreas Sturm

> Home > Persons

[–] **Person information**

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

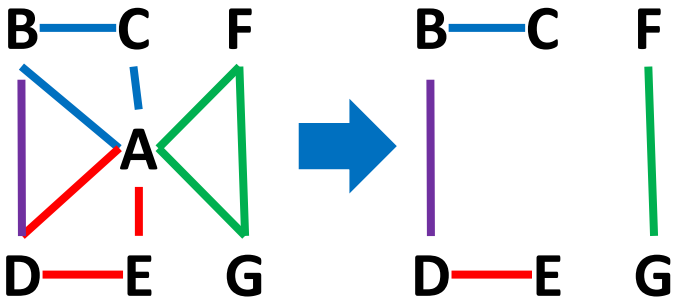
Author disambiguation: Homonyms

Heuristic aproach: **Coautor Graph**

- Nodes are authors of publications
- Edge between authors iff joint publication

Beispiel:

- Paper 1: Autors A, B, C
- Paper 2: Autors A, D, E
- Paper 3: Autors A, F, G
- Paper 4: Autors B, D



How many authors could the profile for author A represent?

- Remove A from coauthor graph
- Every connected subgraph with at least one coauthor of A represents a **coauthor community**
- In the example, we may potentially have two different persons with name A

http://dblp.uni-trier.de/faq/How+does+dblp+detect+coauthor+communities.html

How many authors does a profile represent?

dblp
computer science bibliography

[+] [−] Michael Schumacher

> Home > Persons

2017

[c22] Alevtina Dubovitskaya, Zhigang Xu, Samuel Ryu, Michael Schumacher, Fusheng Wang: **How Blockchain Could Empower eHealth: An Application for Radiation Oncology - (Extended Abstract).** DMAH@VLDB 2017: 3-6

[c21] Davide Calvaresi, Mauro Marinoni, Arnon Sturm, Michael Schumacher, Giorgio C. Buttazzo: **The challenge of real-time multi-agent systems for enabling IoT and CPS.** WI 2017: 356-364

[i1] Alevtina Dubovitskaya, Zhigang Xu, Samuel Ryu, Michael Schumacher, Fusheng Wang: **Secure and Trustable Electronic Medical Records Sharing using Blockchain.** CoRR abs/1709.06528 (2017)

2016

[j4] Sandrine Ding, Michael Schumacher: **Sensor Monitoring of Physical Activity to Improve Glucose Management in Diabetic Patients: A Review.** Sensors 16(4): 589 (2016)

2013

[c20] Bruno Alves, Michael Schumacher, Fabian Cretton, Anne Le Calvé, Gilles Cherix, David Werlen, Christian Gapany, Bertrand Baeryswil, Doris Gerber, Philippe Cloux: **Fairtrace - A Semantic-web Oriented Traceability Solution Applied to the Textile Traceability.** ICEIS (1) 2013: 36-45

[c19] Bruno Alves, Michael Schumacher, Fabian Cretton, Anne Le Calvé, Gilles Cherix, David Werlen, Christian Gapany, Bertrand Baeryswil, Doris Gerber, Philippe Cloux: **Fairtrace: Applying Semantic Web Tools and Techniques to the Textile Traceability.** ICEIS 2013: 68-84

[−] Coauthor Index

1	Bruno Alves	[c20] [c19] [c18] [c16] [c15]
2	Bertrand Baeryswil	[c20] [c19]
3	Federico Bergenti	[c7] [c4]
4	Giorgio C. Buttazzo	[c21]
5	César Cáceres	[c7]
6	Davide Calvaresi	[c21]
7	Anne Le Calvé	[c20] [c19]
8	Fabrice Chantemargue	[c3] [c2] [c1]
9	Gilles Cherix	[c20] [c19]
10	Philippe Cloux	[c20] [c19]
11	Ion Constantinescu	[c13] [c10] [c9] [c8]
12	Fabian Cretton	[c20] [c19] [c16]
13	Thierry Dagaeff	[c1]
14	Sandrine Ding	[j4] [c14]
15	Alevtina Dubovitskaya	[c22] [i1]
16	Boi Faltings	[c10] [c9] [c8]
17	Alberto Fernández 0002	[c7] [c4]
18	Nadine Fröhlich	[c7]
19	Christian Gapany	[c20] [c19]
20	Doris Gerber	[c20] [c19]
21	David Godel	[c15] [c14]
22	Laurent Grangier	[c12] [c11]
23	Heikki Helin	[c7] [c6] [c4]
24	Alexander Helleboogh	[j2]
25	Béat Hirsbrunner	[c3] [c2] [c1]
26	Tom Holvoet	[j2] [j1]
27	Chih-Cheng Hung	[e2]
28	Radu Jurca	[c12] [c11]
29	Oliver Keller	[c7]
30	Ari Kinnunen	[c7] [c4]
31	Matthias Klusch	[c7] [c4]
32	Oliver Krone	[c2] [c1]
33	Heimo Laamanen	[c7]
34	António Lopes	[c7] [c4]
35	Mauro Marinoni	[c21]
36	Andreas Meier 0001	[e1]
37	Francois Mooser	[c14]
38	Alain Mowat	[c13]
39	Henning Müller	[c15] [c14]
40	Abu Khaled Omar	[c15] [c14]
41	Sascha Ossowski	[e2] [c7] [c5]
42	Mathew J. Palakal	[e2]
43	Tim Van Pelt	[c10] [c9] [c8] [c6]

Universität Trier

Author Disambiguation: Synonyms

The same person with different names in different profiles

[+] [–] 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 Bissia
Graphene: A New Protocol for Block Propag
DPM/CBT@ESORICS 2017: 420-428

[+] [–] 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 Identificatio
Comput. 14(6): 620-632

[c8] A. Pinar Ozisik, Gavin A
Graphene: A New Pro
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

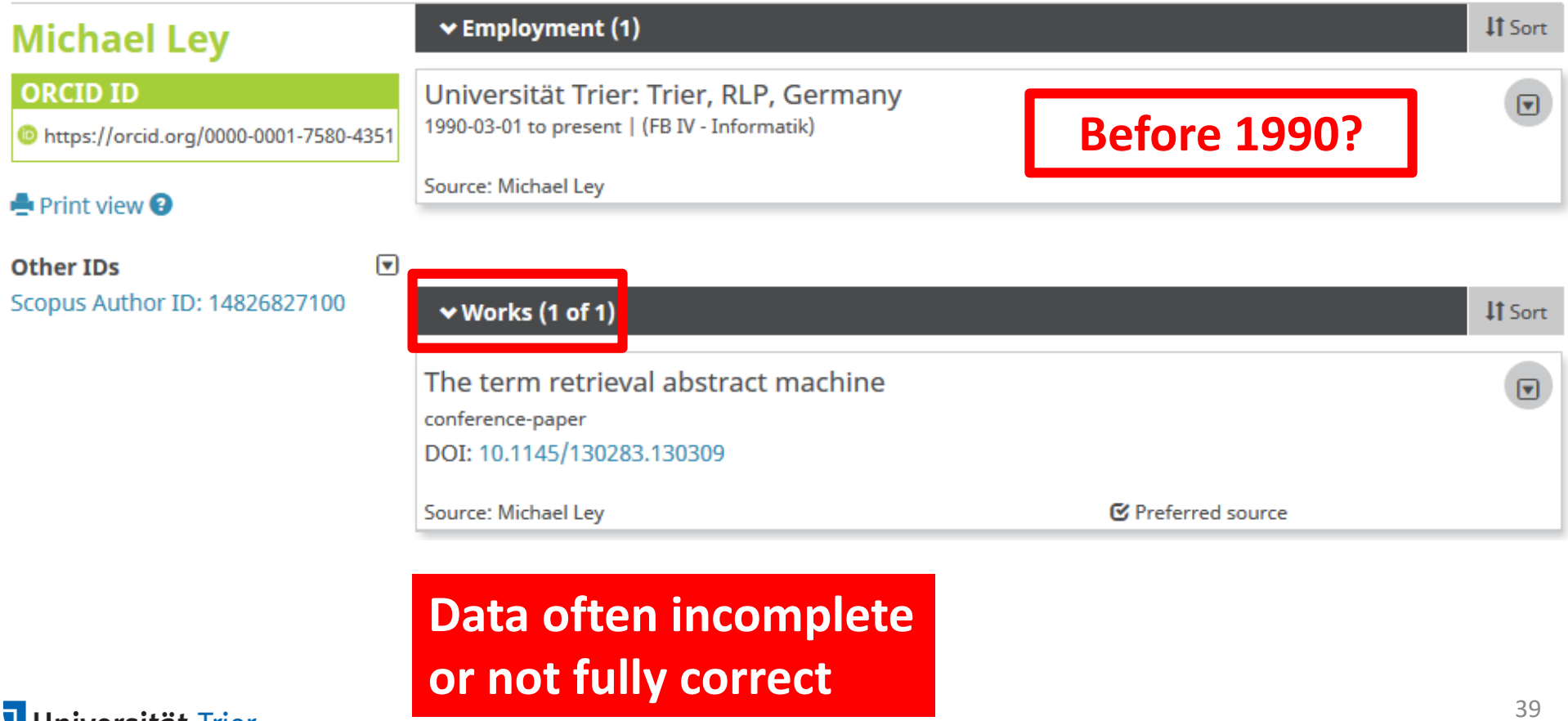
- [PDF] **Bobtail: A Proof-of-Work Target that Minimizes Blockchain Mining Variance**, by George Bissas and Brian Levine. *Presented at the 2017 Scaling Bitcoin Workshop, Palo Alto; arXiv preprint arXiv:1709.08750*, November 2017.
- [PDF] **Market-based Security for Distributed Applications**, by George 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

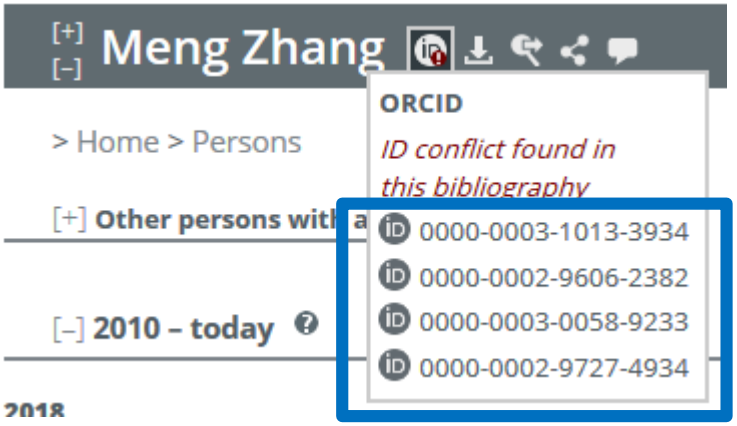
> 5 million ORCIDs

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



Data often incomplete or not fully correct

ORCID for Homonym Detection



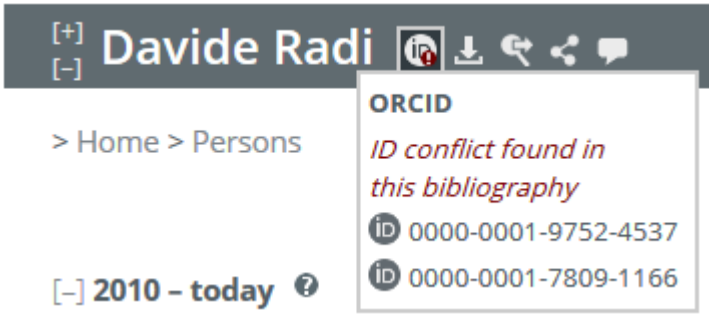
ORCIDs of authors with this name who claimed at least one publication in this profile

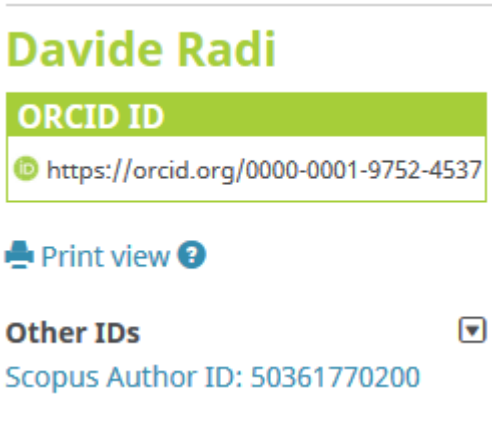
After import of 625,000 ORCIDS:
1,000 candidates for homonyms

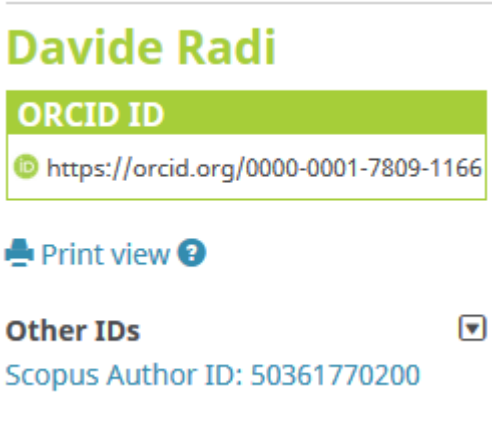
Top candidate:
10 persons in one profile

- Jun Wang 0034 — Xidian University, Institute of Electronic CA
- 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

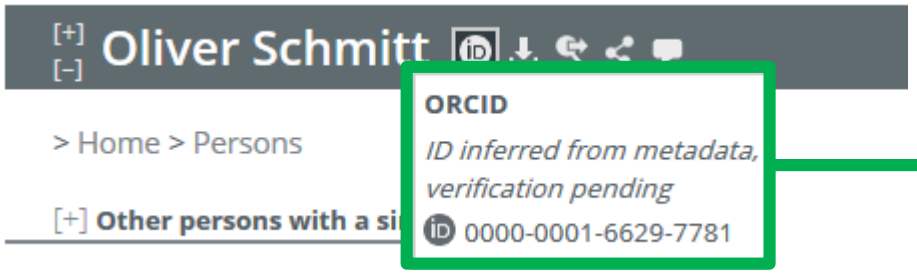
BUT:







ORCID for Synonym Detection



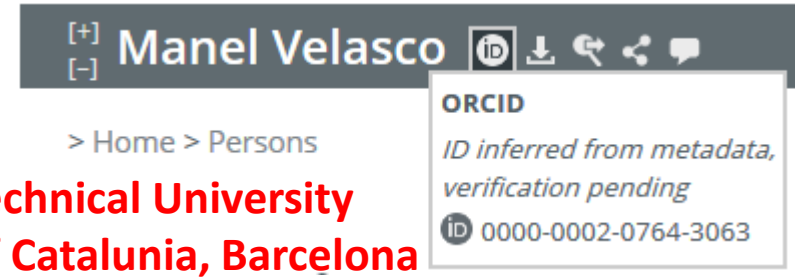
Profiles with common ORCID include papers from the same author (but maybe other papers as well due to homonyms)

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:



Technical University of Catalunia, Barcelona

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

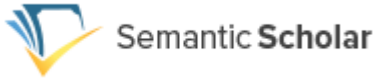
- Author affiliations → better disambiguation, better search
- Keywords
- Topics
- Abstracts
- Full texts

  → better search
- Incoming and outgoing citations
- Performance indicators

  → better result ranking, better conference selection
- …

Sources for Bibliographic Metadata

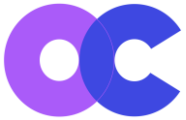
- Dblp.org 

- 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

http://opencitations.net/index/coci

- **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
- COCI, the OpenCitations Index of Crossref open DOI-to-DOI references:
  - 316,243,802 citations
  - 45,145,889 bibliographic resources

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"},
…]

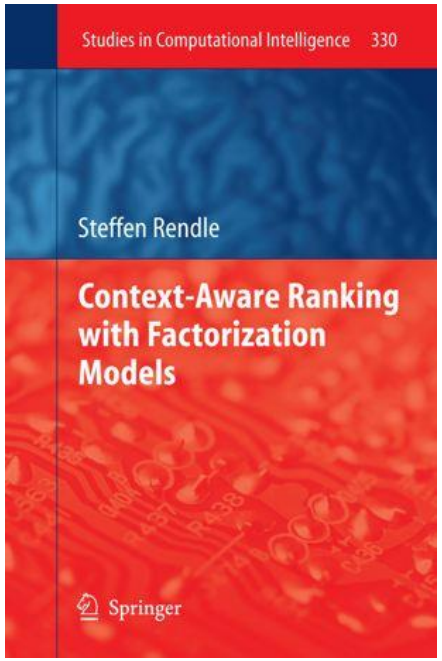
http://api.crossref.org/works/10.1007/978-3-642-25073-6_38

CrossRef Example

23. Rendle, S., Gantner, Z., Freudenthaler, C., Schmidt-Thieme, L.: Fast context-aware recommendations with factorization machines. In: SIGIR (2011)

```
t context-aware recommendations with factorization machines. In  SIGIR (2011)"
```

https://doi.org/10.1007/978-3-642-16898-7



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

Strange names, not linked to a profile

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 ...

☆ 99 Zitiert von: 1 Ähnliche Artikel Alle 6 Versionen »

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

Requires data cleaning

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

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

Semantic Scholar

All Fields ▾ Predicting Query Performance by Query-Drift Estimation

Predicting Query Performance by Query-Drift Estimation

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

Predicting query performance, 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)

17 39 View on ACM Similar Papers More...

TOIS?

Abstract

Predicting query performance, 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

Semantic Scholar

All Fields ▼ Predicting Query Performance by Query-Drift Estimation

Predicting Query Performance by Query-Drift Estimation

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

Predicting **query performance**, 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)

17 39 View on ACM Similar Papers More...

Cited By

Showing 1-10 of 70 extracted citations

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

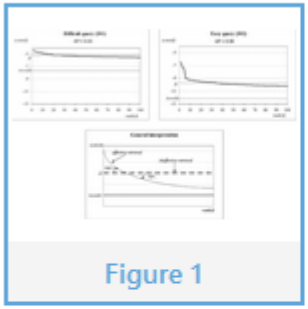


Figure 1

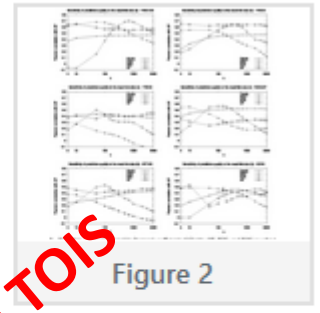


Figure 2

from TOIS

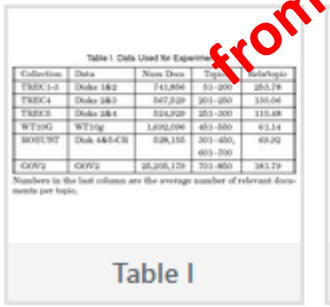


Table I

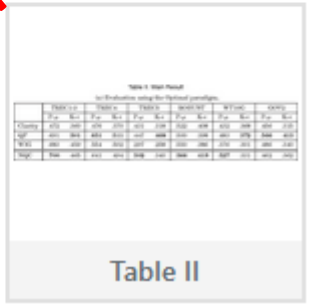
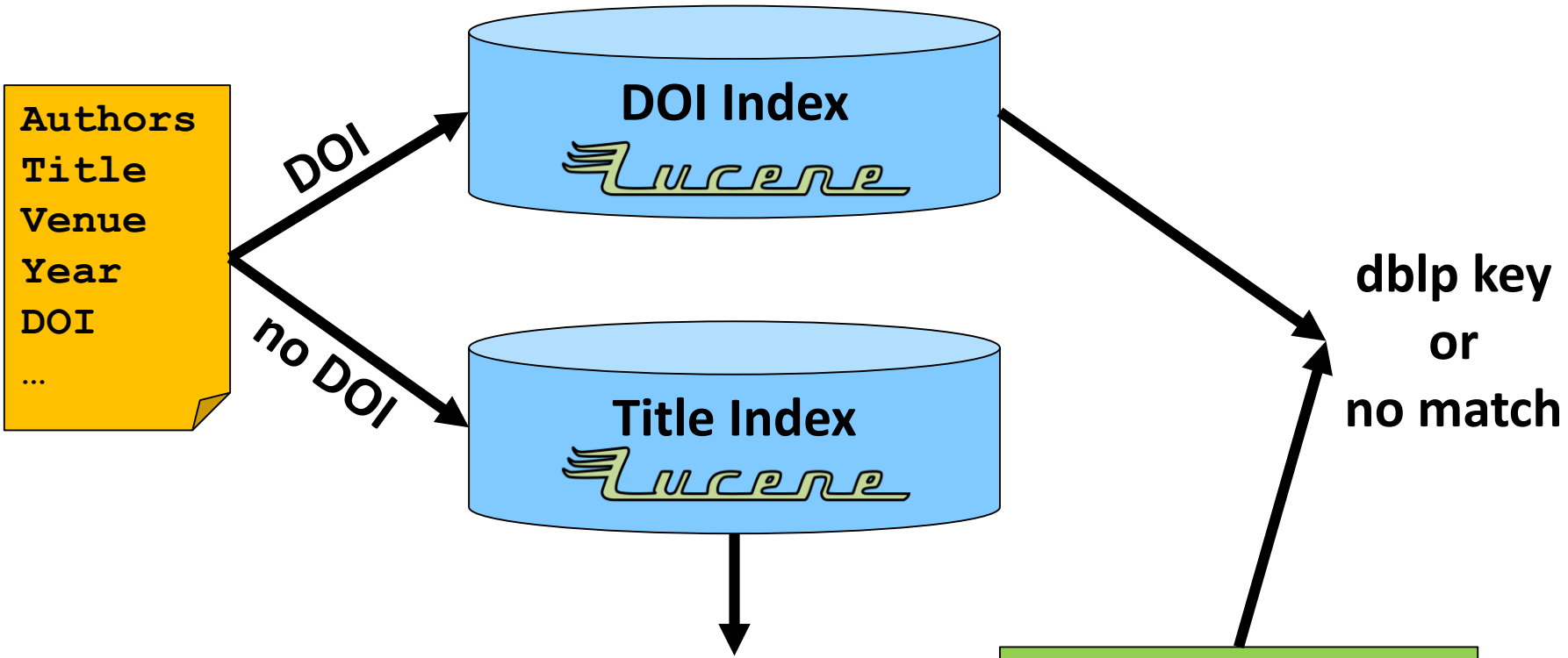


Table II

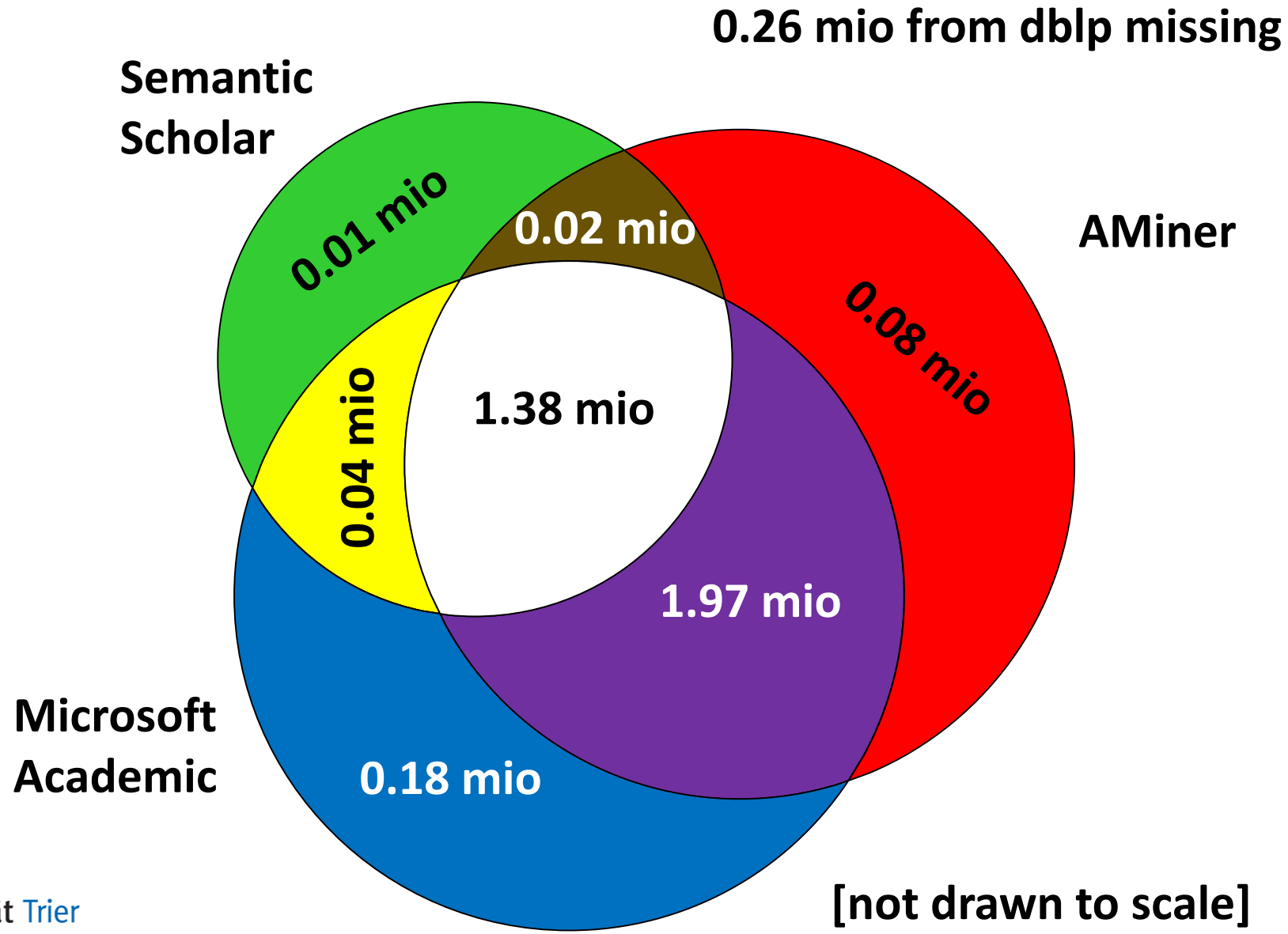
Towards Quantifying Coverage: Mapping papers & citations to dblp

Preprocessing: Index all dblp entries in Lucene



Mapping quality in general very good

Coverage of dblp and Overlap



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

dblp hat retweetet

Jodi Schneider @jschneider · 26. März

#Computerscience is losing the race on @opencitations: 50% of all citations are #CC0 but less than 15% of @dblp_org is covered (And only 20% if you limit to the 2.5 million DBLP publications with @CrossrefOrg metadata.) ceur-ws.org/Vol-2080/paper... #BIR2018 #ECIR2018

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**

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.

Dear Ralf Schenkel
Warm greetings!

… or if it is „fake science“

The paper with the title *News* published in *Datenbank-Spektrum* has impressed us deeply.
Scholars and researchers specializing in a wide range of disciplines have showed interests in your paper.

- 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

Meta data

The *Odyssey* Approach for Optimizing Federated SPARQL Queries

Gabriela Montoya[1(✉)], Hala Skaf-Molli[2], and Katja Hose[1]

[1] Aalborg University, Aalborg, Denmark
{gmontoya,khose}@cs.aau.dk
[2] Nantes University, Nantes, France

```
"name" : "204.pdf",
  "metadata" : {
    "source" : "CRF
    "title" : "The
              SPAR
    "authors" : [ "
                "J
    "emails" : [ "gm
               "ha
    "sections" : [
      "heading" : "I
      "text" : "Fede
```

Grobid 0.5.1 output

```
<analytic>
    <author>
        <persName
            xmlns="http://www.tei-c.org/ns/1.0">
            <forename type="first">Gabriela</forename>
            <surname>Montoya</surname>
        </persName>
        <affiliation key="aff0">
            <orgName type="institution">Aalborg University</orgName>
            <address>
                <country key="DK">Denmark</country>
            </address>
        </affiliation>
    </author>
    <author>
        <persName
            xmlns="http://www.tei-c.org/ns/1.0">
            <forename type="first">Hala</forename>
            <surname>Skaf-Molli</surname>
```

1 Introduction

Federated SPARQL query engines [1,4,7,14,17]
a federation of SPARQL endpoints. Query optir
plex and challenging task in a federated setting.
processing and communication costs by selecti
query. It decomposes the query into subqueries, and produces a query execu-

Citations

References

1. M. Acosta, M. Vidal, T. Lampo, J. C
   Query Processing Engine for SPARQL
2. K. Alexander, R. Cyganiak, M. Hause
   LDOW'09, 2009.

```
"references" : [ {
    "title" : "ANAPSID: An Adaptive Q
    "author" : [ "M. Acosta", "M. Vid
    "venue" : "In ISWC'11,",
    "citeRegEx" : "1",
    "shortCiteRegEx" : "1",
    "year" : 2011
  }, {
    "title" : "Describing Linked Data
    "author" : [ "K. Alexander", "R. C
    "venue" : "In LDOW'09,",
    "citeRegEx" : "2",
    "shortCiteRegEx" : "2",
    "year" : 2009
  }, …
```

Grobid 0.5.1 output

```
            </persName>
        </author>
        <author>
            <persName
                xmlns="http://www.tei-c.org/ns/1.0">
                <forename type="first">E</forename>
                <surname>Ruckhaus</surname>
            </persName>
        </author>
    </analytic>
    <monogr>
        <title level="m">ISWC&amp;apos;11</title>
        <imprint>
            <date type="published" when="2011" />
            <biblScope unit="page" from="18" to="34" />
        </imprint>
    </monogr>
</biblStruct>
<biblStruct xml:id="b1">
    <analytic>
        <title level="a" type="main">Describing Linked Datasets</title>
        <author>
            <persName
                xmlns="http://www.tei-c.org/ns/1.0">
                <forename type="first">K</forename>
                <surname>Alexander</surname>
            </persName>
        </author>
        <author>
            <persName
                xmlns="http://www.tei-c.org/ns/1.0">
                <forename type="first">R</forename>
                <surname>Cyganiak</surname>
            </persName>
```

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
    }, …
```

Fulltext Data Set

- Manual collection of available PDFs from DVDs, Web sites, …
  - (Almost) complete set of publications for DB/IR: SIGMOD, SIGIR, (P)VLDB, EDBT, TOIS, TODS, IR, TKDE, CACM, …
  - ACL Anthology
  - WWW, IJCAI, ISWC, CoRR, …
- Semi-automatic mapping to dblp keys
- Extraction of full text, citations, citation ctxs

~170.000 documents available (plus ~30.000 waiting to be processed)

Evaluating Mapping Quality for Citations

96 papers from PVLDB Volume 10, converted with ScienceParse

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

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

- Recall: ~96%
- Precision: ~97.5%

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

Evaluating Mapping Quality for Citations

Try to re-find references from Crossref already mapped to DOIs

{"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"}

DOI match

conf/esws/ArandaAC11

Parse with citation parser & textual match

journals/ws/ArandaACP13

Evaluating Mapping Quality for Citations

- 2,516 articles from SIGIR, ECIR, SIGMOD, ISWC, TOIS, IR Journal, VLDB Journal from 2013 or newer with 29,932 matching citations in Crossref
- Citation parsing with **Cermine**:
  22282 correctly matched (0.744%),
  160 incorrect matches, 7283 not matched
- Citation parsing with **Grobid**:
  27993 correctly matched (0.935%),
  317 incorrect matches, 1125 not matched

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
[illegible]	[illegible]	[illegible]
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	53
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)