SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

Universität Trier

Autoren-Disambiguierung in bibliographischen Metadaten

Der Fall dblp

Marcel R. Ackermann

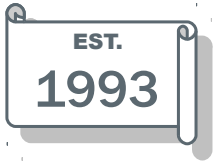
4. April 2016

Was ist dblp?

Indexierung von Publikationen

3,3 Millionen Publikationen aus der Informatik und Nachbardisziplinen

Was ist dblp?



Indexierung von Publikationen

home | browse | search | about

dbl
compute

[+] [–] Search dblp

> Home

[–] Venue results

Likely matches

- International Journal of
- International Conferenc
- Metadata Management

[–] Publication results

found 3,375 matches

2016

Gregory Gi
Nicolas Ray
Facilitatin
metadata
Earth Obse
(2016)

Evangelos
Emiliano Pe

3,3 Million
Informatik

Profile von Konferenzen und Journalreihen

home | browse | search | about

dblp
computer science bibliography

search dblp

ACM/IEEE Joint Conference on Digital Libraries (JCDL)

> Home > Conferences and Workshops

modern | Dagstuhl

www.jcdl.org

ACM DL - IEEE ADL

15. JCDL 2015: Knoxville, TN, USA

Paul Logasa Bogen II, Suzie Allard, Holly Mercer, Micah Beck, Sally Jo Cunningham, Dion Hoe-Lian Goh, Geneva Henry:
Proceedings of the 15th ACM/IEEE-CE Joint Conference on Digital Libraries, Knoxville, TN, USA, June 21-25, 2015. ACM 2015, ISBN 978-1-4503-3594-2 [contents]

Ben Fields, Kevin R. Page:
Proceedings of the 2nd International Workshop on Digital Libraries for Musicology, DLfM@JCDL 2015, Knoxville, TN, USA, June 25, 2015. ACM 2015, ISBN 978-1-4503-3563-8 [contents]

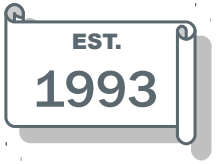
14. JCDL 2014: London, UK

IEEE/ACM Joint Conference on Digital Libraries, JCDL 2014, London, United Kingdom, September 8-12, 2014. IEEE 2014, ISBN 978-1-4799-5569-5 [contents]

eamus Ross:
n Semantic Digital Archives (SDA 2014)
s Conference (DL 2014), London, UK,
1306, CEUR-WS.org 2014 [contents]

69000 Inhaltsverzeichnisse
4650 Konferenzreihen
1450 Journale

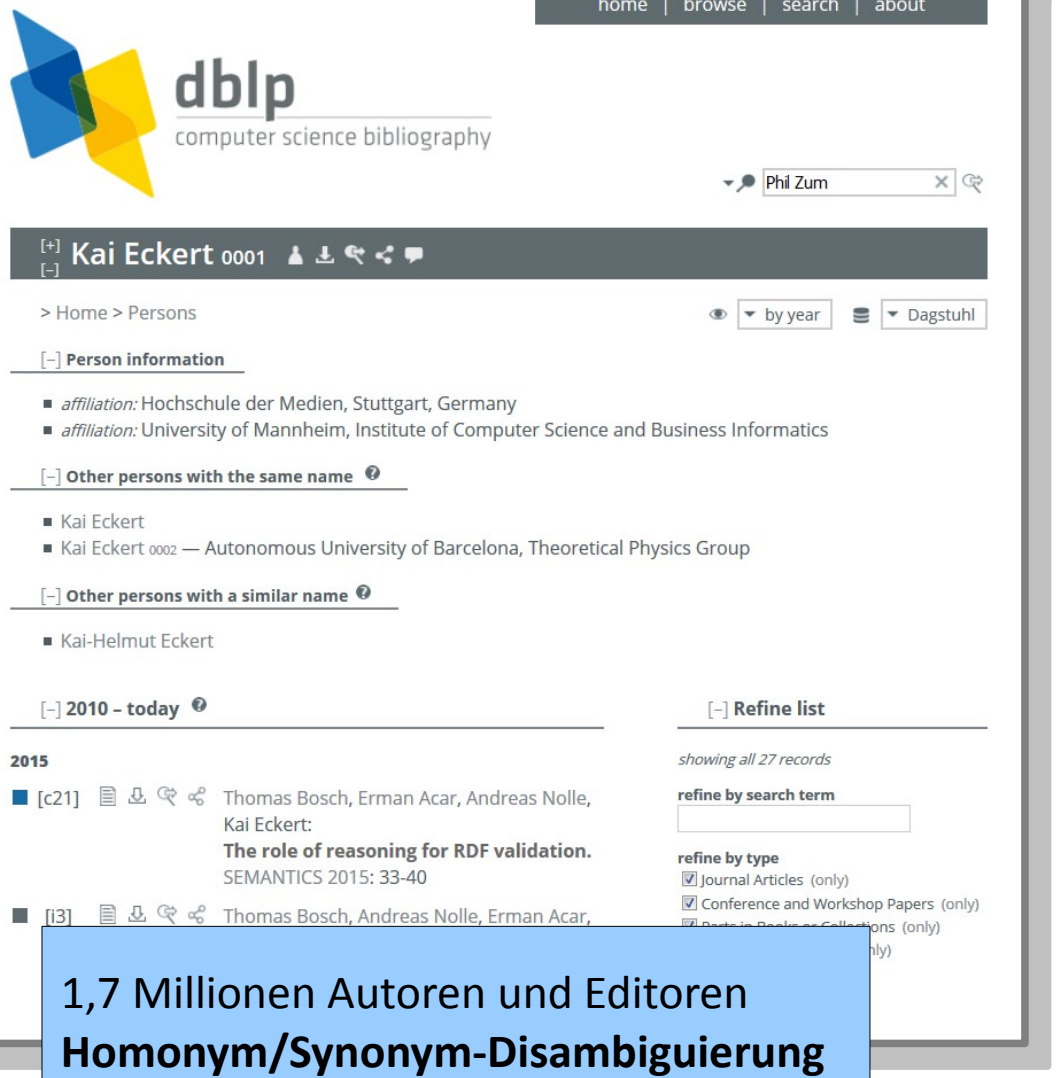
Was ist dblp?



Indexierung von Publikationen

Profile von Konferenzen und Journalreihen

Autoren-Profile



3,3 Million
Informatik

69000 Inhal
4650 Konfe
1450 Journa

1,7 Millionen Autoren und Editoren
Homonym/Synonym-Disambiguierung

Was ist dblp?

EST. 1993

Indexierung von Publikationen

home | browse | search | about

[+] [–] Search dblp

> Home

[–] Venue results

Likely matches

- International Journal of
- International Conferenc
- Metadata Management

[–] Publication results

found 3,375 matches

2016

Gregory Gi
Nicolas Ray
Facilitatin
metadata
Earth Obse
(2016)

Evangelos
Emiliano Pe

3,3 Million
Informatik

Profile von Konferenzen und Journalreihen

home | browse | search | about

ACM/IEEE Joint

> Home > Conferences and

www.jcdl.org

ACM DL - IEEE ADL

15. JCDL 2015: Knoxville, T

Paul Logasa
Hoe-Lian Goh
Proceedings
USA, June 21

Ben Fields, K
Proceedings
DLfM@JCDL
[contents]

14. JCDL 2014: London, UK

IEEE/ACM Joi
September 8

69000 Inhal
4650 Konfe
1450 Journa

Autoren-Profile

home | browse | search | about

[+] [–] Kai Eckert 0001

> Home > Persons

[–] Person information

- *affiliation:* Hochschule der
- *affiliation:* University of M

[–] Other persons with the sa

- Kai Eckert
- Kai Eckert 0002 — Autonor

[–] Other persons with a sim

- Kai-Helmut Eckert

[–] 2010 – today

2015

[c21] Thom
Kai Ec
The r
SEMA

[i3] Thom

1,7 Million
Homonym/Synonym-Disambiguierung

Daten-APIs

```
– <rdf:RDF>
  – <dblp:Person rdf:about="http://dblp.dagstuhl.de/pers/b/Bizer:Christian">
      <owl:sameAs rdf:resource="http://dblp.org/pers/b/Bizer:Christian"/>
      <owl:sameAs rdf:resource="http://scholar.google.com/citations?user=INQwsQkAAAAJ"/>
      <owl:sameAs rdf:resource="http://dl.acm.org/author_page.cfm?id=81100002809"/>
      <dblp:personLastModifiedDate>2016-03-06T00:20:58+0100</dblp:personLastModifiedDate>
      <dblp:primaryFullPersonName>Christian Bizer</dblp:primaryFullPersonName>
    – <dblp:primaryAffiliation>
        University of Mannheim, School of Business Informatics and Mathematics
      </dblp:primaryAffiliation>
    – <dblp:otherAffiliation>
        Free University of Berlin, Department of Information Systems
      </dblp:otherAffiliation>
      <dblp:primaryHomepage rdf:resource="http://dws.informatik.uni-mannheim.de/en/people/professors/prof-dr-christian-bizer/"/>
      <dblp:otherHomepage rdf:resource="http://www.wiwiss.fu-berlin.de/en/institute/pwo/bizer/team/BizerChristian.html"/>
      <dblp:authorOf rdf:resource="http://dblp.dagstuhl.de/rec/journals/semweb/LehmannIJJKMHMK15"/>
      <dblp:authorOf rdf:resource="http://dblp.dagstuhl.de/rec/journals/ws/RistoskiBP15"/>
      <dblp:authorOf rdf:resource="http://dblp.dagstuhl.de/rec/journals/ws/LehmbergRRMPB15"/>
      <dblp:authorOf rdf:resource="http://dblp.dagstuhl.de/rec/conf/ecweb/MeuselPMPB15"/>
      <dblp:authorOf rdf:resource="http://dblp.dagstuhl.de/rec/conf/wims/RitzeLB15"/>
      <dblp:authorOf rdf:resource="http://dblp.dagstuhl.de/rec/conf/wims/MeuselBP15"/>
      <dblp:authorOf rdf:resource="http://dblp.dagstuhl.de/rec/conf/www/AuerBBH15"/>
      <dblp:authorOf rdf:resource="http://dblp.dagstuhl.de/rec/conf/www/MeuselSBP15"/>
      <dblp:editorOf rdf:resource="http://dblp.dagstuhl.de/rec/conf/www/2015ldow"/>
      <dblp:authorOf rdf:resource="http://dblp.dagstuhl.de/rec/journals/ijswis/PaulheimB14"/>
      <dblp:authorOf rdf:resour
      <dblp:authorOf rdf:resour
      <dblp:authorOf rdf:resour
      <dblp:authorOf rdf:resour
      <dblp:authorOf rdf:resour
      <dblp:authorOf rdf:resour
      <dblp:authorOf rdf:resour
      <dblp:authorOf rdf:resour
      <dblp:authorOf rdf:resour
      <dblp:authorOf rdf:resour
      <dblp:authorOf rdf:resour
      <dblp:authorOf rdf:resour
      <dblp:authorOf rdf:resour
      <dblp:authorOf rdf:resource="http://dblp.dagstuhl.de/rec/series/lncs/BrylBIVHJYC14"/>
      <dblp:editorOf rdf:resource="http://dblp.dagstuhl.de/rec/conf/www/2014ldow"/>
```

offene Datenschnittstellen,
kompletter Datensatz als
Open-Data unter ODC-by Lizenz

dblp in Überblick

Grundsätze

- Offenheit
- Neutralität der Daten
- **Datenqualität**
- Semantische Anreicherung
- Ermöglichen von Forschung
- Nutzerorientierung

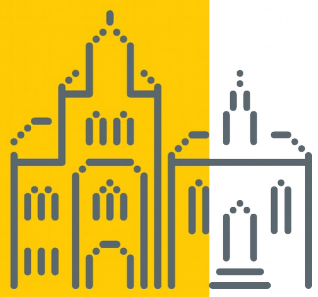
dblp in Überblick

Grundsätze

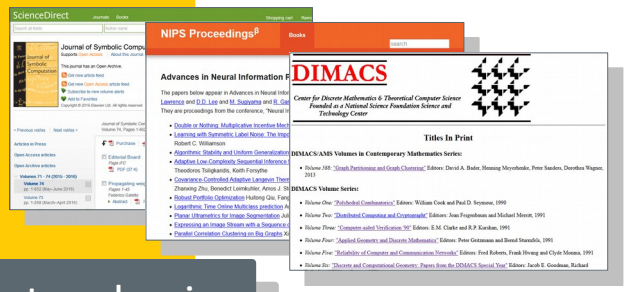
- Offenheit
- Neutralität der Daten
- **Datenqualität**
- Semantische Anreicherung
- Ermöglichen von Forschung
- Nutzerorientierung

Was dblp nicht ist ...

- keine Digital Library
- kein bibliothekarisches Projekt
- kein fachübergreifender Dienst
- „kein Geschäftsmodell“
- kein großes Team



Der dblp-Workflow



1. Datenakquise

- Webcrawler oder Datenlieferungen
- Normalisierung

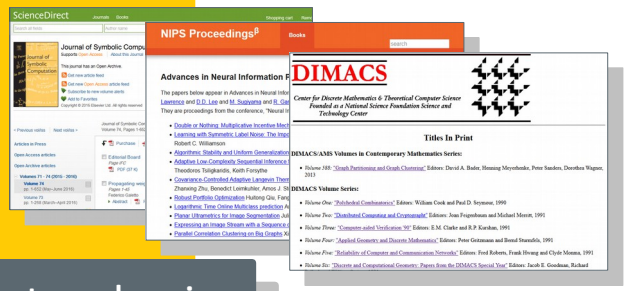
Publisher

Internes Datenformat (XML)

```
<article key="journals/jsc/Kemper16">
<author>Gregor Kemper</author>
<title>Using extended Derksen ideals in computational
invariant theory.</title>
<pages>161-181</pages>
<year>2016</year>
<volume>72</volume>
<journal>J. Symb. Comput.</journal>
<ee>http://dx.doi.org/10.1016/j.jsc.2015.02.004</ee>
<url>db/journals/jsc/jsc72.html#Kemper16</url>
</article>
```

dblp

Der dblp-Workflow



1. Datenakquise

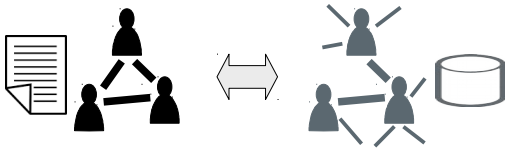
- Webcrawler oder Datenlieferungen
- Normalisierung

Publisher

Internes D

```
<article key="jour
<author>Gregor Kem
<title>Using exter
invariant theory.<
<pages>161-181</pa
<year>2016</year>
<volume>72</volume
<journal>J. Symb.
<ee>http://dx.doi.
<url>db/journals/j
</article>
```

2. Initiale Datensäuberung

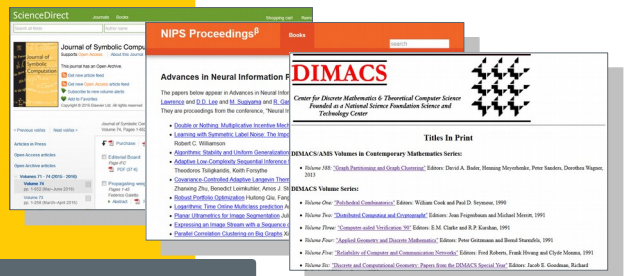


- **Zuordnen der Autorenschaft**
- **Homonym/Synonym-Erkennung**
- Korrektur der Neudaten
- Vervollständigungen (zB: abgek. Namen, fehlende Namensteile)
- Anreicherung (Homepages, Affiliations, PIDs, zB ORCIDs, ...)

eigener Korpus als Authority

dblp

Der dblp-Workflow



1. Datenakquise

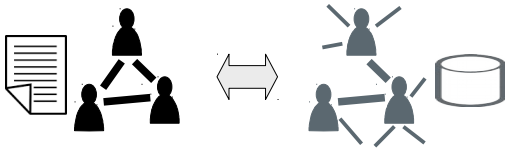
- Webcrawler oder Datenlieferungen
- Normalisierung

Publisher

Internes D

```
<article key="jour
<author>Gregor Kem
<title>Using exter
invariant theory.<
<pages>161-181</pa
<year>2016</year>
<volume>72</volume
<journal>J. Symb.
<ee>http://dx.doi.
<url>db/journals/j
</article>
```

2. Initiale Datensäuberung

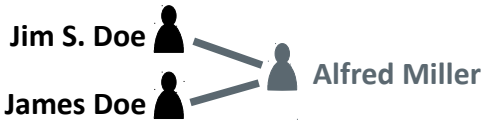


- **Zuordnen der Autorenschaft**
- **Homonym/Synonym-Erkennung**
- Korrektur der Neudaten
- Vervollständigungen (zB: abgek. Namen, fehlende Namensteile)
- Anreicherung (Homepages, Affiliations, PIDs, zB ORCIDs, ...)

eigener Korpus als Authority

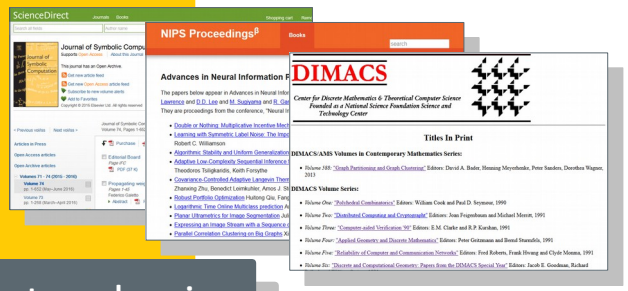
Netzwerk-analysen

3. Kontinuierliche Qualitätskontrolle



- **Homonym/Synonym-Erkennung**
- Korrekturen am bestehenden Corpus
- neue Informationen durch neuste Ergänzungen

Der dblp-Workflow



1. Datenakquise

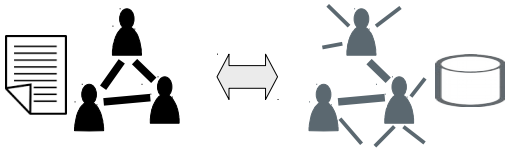
- Webcrawler oder Datenlieferungen
- Normalisierung

Internes D

```
<article key="jour
<author>Gregor Kem
<title>Using exter
invariant theory.<
<pages>161-181</pa
<year>2016</year>
<volume>72</volume
<journal>J. Symb.
<ee>http://dx.doi.
<url>db/journals/j
</article>
```

Publisher

2. Initiale Datensäuberung



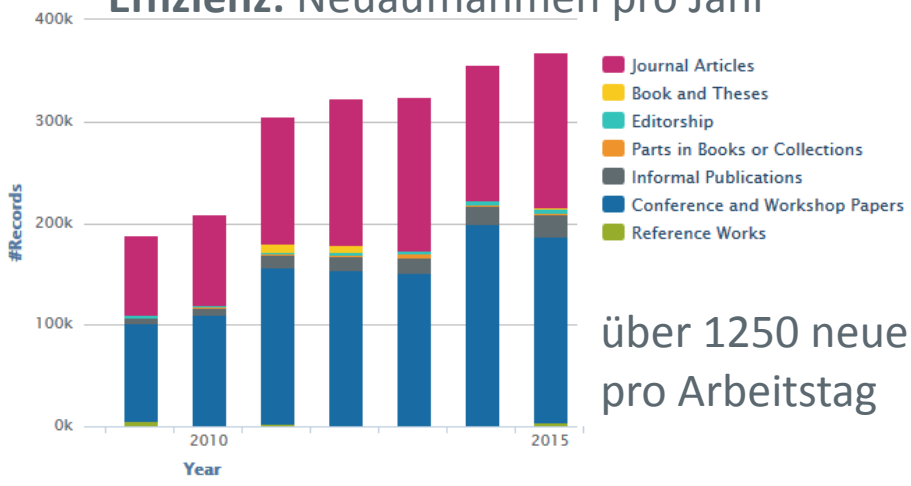
- **Zuordnen der Autorenschaft**
- **Homonym/Synonym-Erkennung**
- Korrektur der Neudaten
- Vervollständigungen (zB: abgek. Namen, fehlende Namensteile)
- Anreicherung (Homepages, Affiliations, PIDs, zB ORCIDs, ...)

eigener Korpus als Authority

dblp

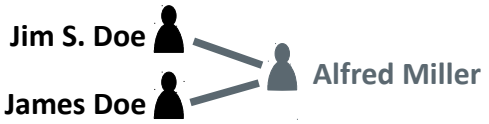
Netzwerk-analysen

Effizienz: Neuaufnahmen pro Jahr

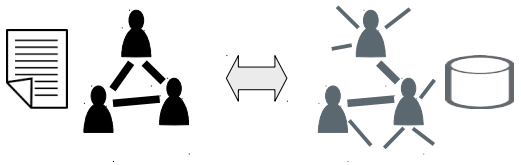


über 1250 neue Artikel pro Arbeitstag

3. Kontinuierliche Qualitätskontrolle



- **Homonym/Synonym-Erkennung**
- Korrekturen am bestehenden Corpus
- neue Informationen durch neuste Ergänzungen

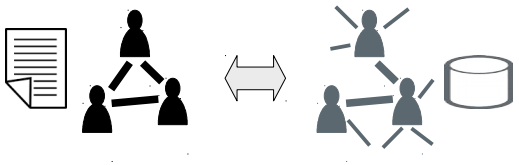


Disclaimer:

Your mileage may vary!

- optimiert für die Informatik
- basiert nur auf Kern-Metadaten

Algorithmus *dblpi*



1. Initiale Kandidatenmenge erzeugen

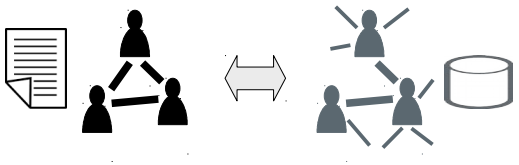
- Suche auf gesamtem Korpus (rein syntaktisch)
- benötigt sehr schnelle Ähnlichkeitsmaße und Datenstrukturen

Simple Blocking-Strategien, z.B.:

- letzter Namensteil
- längster Namensteil
- seltenster Namensteil
- vorkommende Konsonanten

alle Vergleiche unter Unifikation von diakritischen Zeichen

Algorithmus *dblpi*



1. Initiale Kandidatenmenge erzeugen

- S
- b

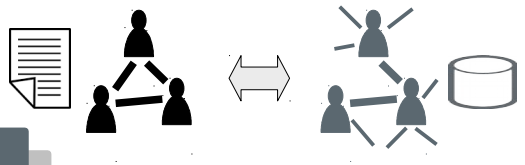
2. Zu große Kandidatenmengen adaptiv filtern

- zu groß: mehr als 10 Kandidaten (z.B. „W. Wang“ ≈ 15000)
- Filter darf teurere Maße verwenden als initiale Erstellung

Mögliche Filterregel:

- alle Namensteile als Präfix enthalten
- in eine oder andere Richtung
- ggf. permutierte Reihenfolge

Algorithmus *dblpi*



1. Initiale Kandidatenmenge erzeugen

- S
- b

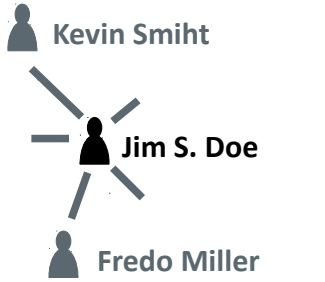
2. Zu große Kandidatenmengen adaptiv filtern

- zu
- Filt

3. Weitere Kandidaten im Kontext finden

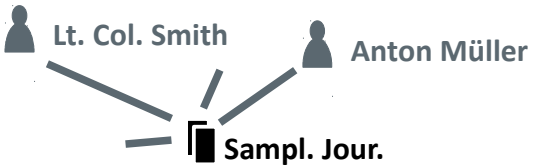
- Ausnutzen von semantischen Zusammenhängen
- darf teurere Ähnlichkeitsmaße verwenden

Kontext der Kandidatenmenge:

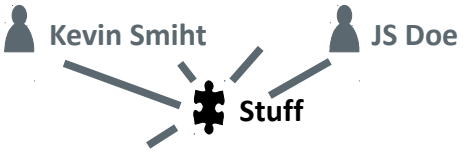


Co-Autoren-Netzwerk

weitere Kontexte, z.B.:



Reihen-Autoren-Netzwerk



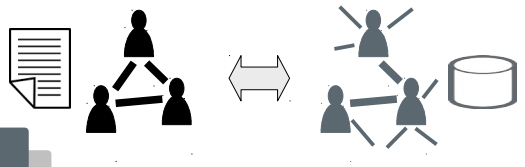
Themen-Autoren-Netzwerk

Mögliche Maße:

- Levenshtein
- Jaro-Winkler
- n-Gramme
- Soundex-Varianten

mit Anpassungen an Personennamen

Algorithmus *dblpi*



1. Initiale Kandidatenmenge erzeugen

- S
- b

2. Zu große Kandidatenmengen adaptiv filtern

- zu
- Filt

3. Weitere Kandidaten im Kontext finden

- Ausn
- darf

Konte
Kand

Co-A

Themen-Autoren-Netzwerk

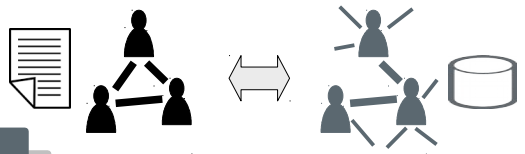
4. Kandidaten bewerten

- Bewertung anhand mehrdimensionaler Features
- Werte je zwischen 0 und 1

ähnlicher Name	publiziert in gleicher Reihe
gemeinsame Publikation	publiziert in *ähnlicher* Reihe
zeitnahe Publikationen	verwendet gleiche Themen
gleiche Affiliation	ggf. weitere ...

- Ranking mittels gewichteter Summe

Algorithmus *dblpi*



1. Initiale Kandidatenmenge erzeugen

- S
- b

2. Zu große Kandidatenmengen adaptiv filtern

- zu
- Filt

3. Weitere Kandidaten im Kontext finden

- Ausn
- darf

Konte
Kand

Co-A

4. Kandidaten bewerten

- Bewert
- Werte

ähnl

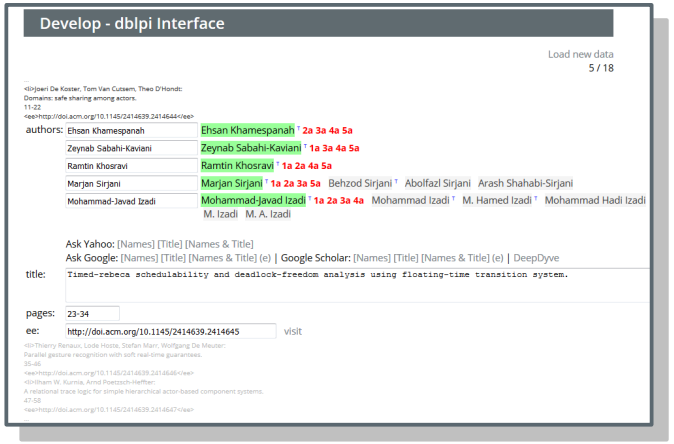
gem

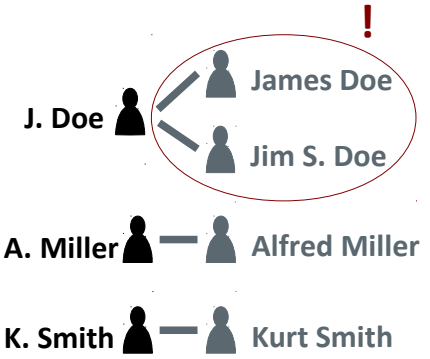
zeitn

gleic

- Rankin

5. Autoren zuordnen

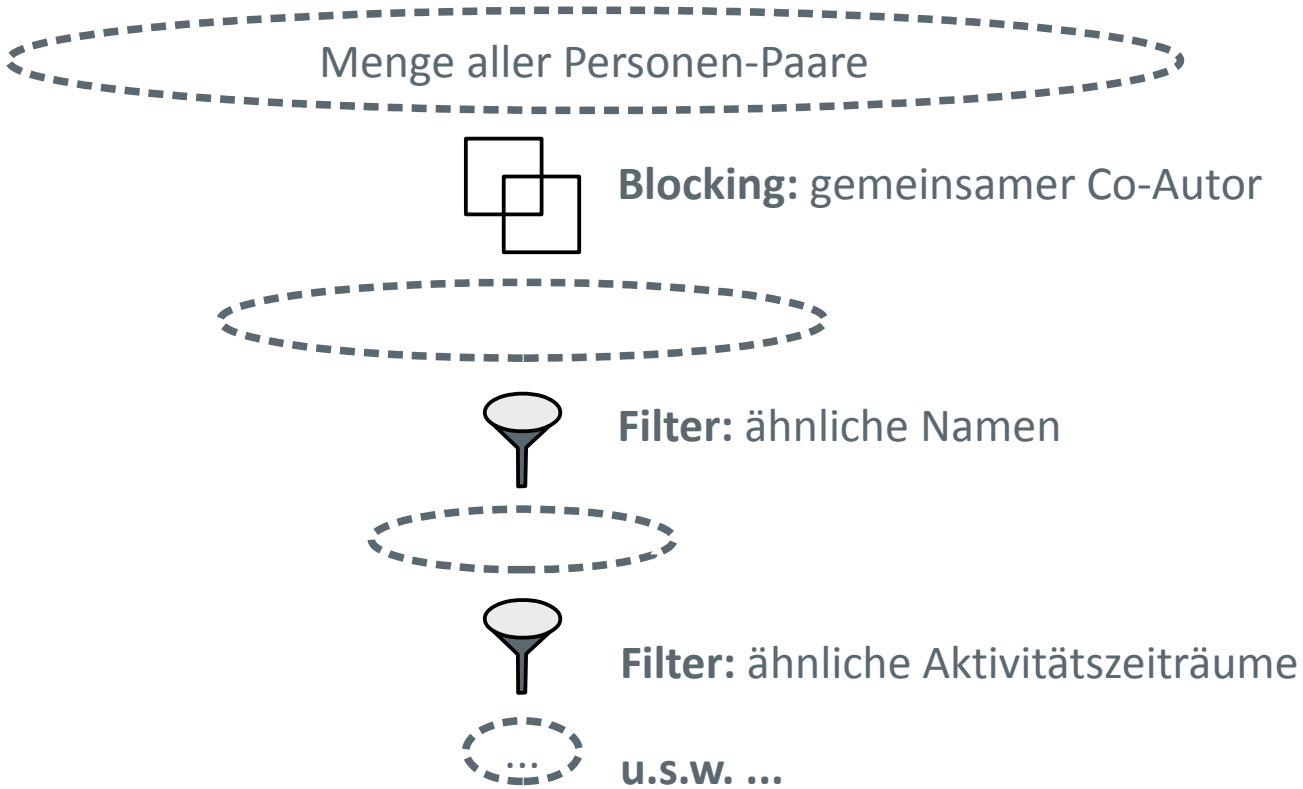




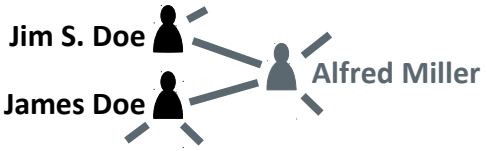
- manuelle Zuordnung anhand der Kandidatenlisten
- im Zweifel: Recherche mit externen Quellen
- Fehler im Korpus werden durch die Vorschläge oft sichtbar!

Kontinuierliche Qualitätskontrolle

Synonyme entdecken

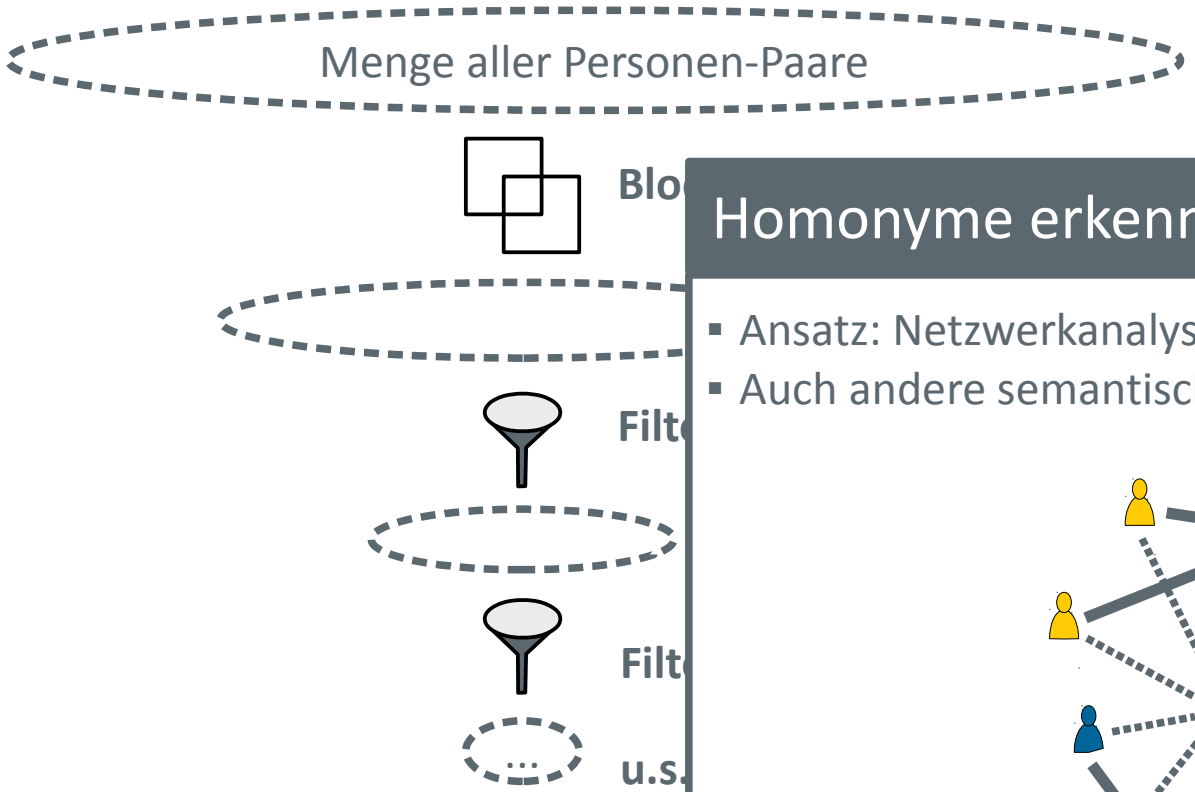


- Ranking der Vorschlagslisten & manuelle Kontrolle

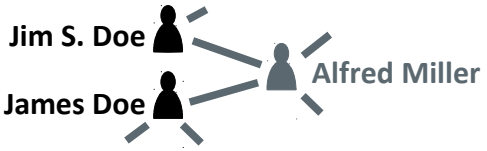


Kontinuierliche Qualitätskontrolle

Synonyme entdecken

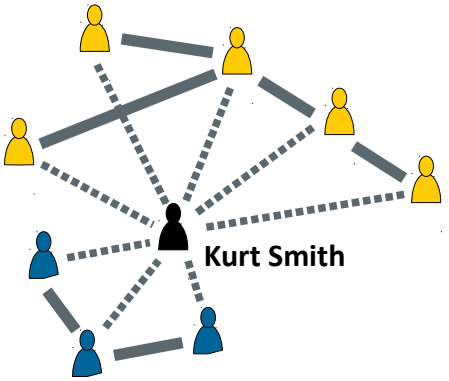


- Ranking der Vorschlagslisten & ma



Homonyme erkennen

- Ansatz: Netzwerkanalyse von Co-Autor-Communities
- Auch andere semantische Netzwerke nutzbar



- Derzeit noch kein automatisiertes Verfahren im Einsatz
- Erforschung solcher Teil von laufendem Drittmittel-Projekt

Vielen Dank!

http://dblp.dagstuhl.de
http://dblp.uni-trier.de
http://dblp.org

✉ **dblp@dagstuhl.de**
@dblp_org