



**dblp**

computer science bibliography

# Data integration at the dblp computer science bibliography

Marcel R. Ackermann

April 15, 2024



# What is dblp? (1/4)

The screenshot shows the dblp website interface. At the top, it says "a service of SCHLOSS DAGSTUHL Leibniz Center for Informatics" with navigation links for "home | browse | search | about". The dblp logo is on the left, and a search bar contains "open data". Below the search bar is a dark banner with "[+] Search dblp for Publications" and "powered by CompleteSearch, courtesy of Hannah Bast, University of Freiburg". The main content area shows "> Home > Search" and a "Publication search results" section with a download icon. It states "found 6,749 matches" and lists results for 2022 and 2021. A "Refine list" section on the right includes a bar chart showing publication counts from 1962 to 2022, peaking at 900 in 2022, and a list of authors to refine by: Sören Auer (19), Sebastian Neumaier (18), Silvio Peroni (18), and 18,978 more options.

SCHLOSS DAGSTUHL  
Leibniz Center for Informatics

home | browse | search | about

dblp  
computer science bibliography

open data

[+] Search dblp for Publications ⓘ  
powered by CompleteSearch, courtesy of Hannah Bast, University of Freiburg

> Home > Search

[+] Publication search results ↓

found 6,749 matches

2022

Yang Yang, Hongchen Wei, Zhen-Qiang Sun, Guang-Yu Li, Yuanchun Zhou, Hui Xiong, Jian Yang:  
**S2OSC: A**  
**Classifica**  
(2022)

2021

Matthias  
**Knowledge-Driven Harmonization of Sensor Observations:  
Exploiting Linked Open Data for IoT Data Streams.** Karlsruhe  
Institute of Technology, Germany, KIT Scientific Publishing 2021

refine by author

1962 2022 900

Sören Auer (19)  
Sebastian Neumaier (18)  
Silvio Peroni (18)  
18,978 more options

Open indexing service and scholarly search engine for 7.1+ million computer science publications



# What is dblp? (2/4)

[+] Silvio Peroni     

> Home > Persons

 by year  Dagstuhl

## [+] Person information

- affiliation: University of Bologna, Department of Classical Philology and Italian Studies,
- affiliation: OpenCitations.net

## [+] 2020 - today

2021

- [j29]     Ivan Heibi , Silvio Peroni :  
**A qualitative and quantitative analysis of open citations to retracted articles: the Wakefield 1998 et al.'s case.** *Scientometrics* 126(10): 8433-8470 (2021)
- [c82]     Zeyd Boukhers, Philipp Mayr, Silvio Peroni:  
**BiblioDAP'21: The 1st Workshop on Bibliographic Data Analysis and Processing.** *KDD* 2021: 4110-4111
- [i25]     Federica Bologna, Angelo Di Iorio, Silvio Peroni, Francesco Poggi:  
**Do open citations inform the qualitative peer-review evaluation in research assessments? An analysis of the Italian National Scientific Qualification.** *CoRR abs/2103.07942* (2021)
- [i24]     Federica Bologna, Angelo Di Iorio, Silvio Peroni, Francesco Poggi:  
**Can we assess research using open scientific knowledge graphs? A case study within the Italian National Scientific Qualification.** *CoRR abs/2105.08599* (2021)

Curated bibliographies  
for 3.4+ million  
authors and editors



refine by search term

refine by type

- Journal Articles (only)
- Conference and Workshop Papers (only)
- Parts in Books or Collections (only)
- Editorship (only)
- Informal Publications (only)

select all | deselect all

refine by coauthor

Fabio Vitali (63)  
Angelo Di Iorio (35)  
David M. Shotton (34)  
Andrea Giovanni Nuzzolese (31)



# What is dblp? (3/4)

## [+] ACM Conference on Human Factors in Computing Systems (CHI) [-]

> Home > Conferences and Workshops

Trier 2









### [-] Description

The **ACM Conference on Human Factors in Computing Systems (CHI)** series of academic conferences is generally considered the most prestigious in the field of human-computer interaction and is one of the top-ranked conferences in computer science. It is hosted by ACM SIGCHI, the Special Interest Group on computer-human interaction. CHI has been held annually since 1982 and attracts thousands of international attendees. CHI 2020, which was originally planned to take place on April, was cancelled due to COVID-19, and CHI 2021 will be held online as a virtual conference chaired by Yoshifumi Kitamura and Aaron Quigley. CHI 2021 "making waves, combining strengths" was originally scheduled to take place in Yokohama.

 read full information at Wikipedia

[Text kindly provided by Wikipedia under CC-BY-SA-3.0 license.]

### CHI 2021: Virtual Event / Yokohama, Japan

-     Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, Steven Mark Drucker:  
**CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021.** ACM 2021, ISBN 978-1-4503-8096-6 [contents] ■ 7.6
-     Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi:  
**CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama Japan, May 8-13, 2021, Extended Abstracts.** ACM 2021, ISBN 978-1-4503-8095-9 [contents] ■ 5.9

### [-] Venue statistics

#### records by year



#### frequent topics

design interact interfac social system  
user visual

(currently stemmed word rumps only)

#### frequent authors

- Stephen A. Brewster (109)
- Hiroshi Ishii <sup>0001</sup> (101)
- Patrick Olivier (98)
- Florian 'Floyd' Mueller (95)
- Albrecht Schmidt <sup>0001</sup> (87)
- Patrick Baudisch (86)

Open directory of  
**1,800+** computer  
science journals and  
**6,600+** conference  
and workshop series












# What is dblp? (4/4)

```
<rdf:RDF>
  <dblp:Person rdf:about="https://dblp.org/pid/48/1099">
    <rdf:type rdf:resource="https://dblp.org/rdf/schema#Creator"/>
    <owl:sameAs rdf:resource="https://orcid.org/0000-0003-0530-4305"/>
    <owl:sameAs rdf:resource="https://www.wikidata.org/entity/Q30536251"/>
    <owl:sameAs rdf:resource="https://id.loc.gov/authorities/names/no2014111866"/>
    <owl:sameAs rdf:resource="https://d-nb.info/gnd/105432011X"/>
    <rdfs:label>Silvio Peroni</rdfs:label>
    <dblp:orcid rdf:resource="https://orcid.org/0000-0003-0530-4305"/>
    <dblp:wikidata rdf:resource="https://www.wikidata.org/entity/Q30536251"/>
    <+<datacite:hasIdentifier></datacite:hasIdentifier>
    <+<datacite:hasIdentifier></datacite:hasIdentifier>
    <+<datacite:hasIdentifier></datacite:hasIdentifier>
    <+<datacite:hasIdentifier></datacite:hasIdentifier>
    <+<datacite:hasIdentifier></datacite:hasIdentifier>
    <+<datacite:hasIdentifier></datacite:hasIdentifier>
    <+<datacite:hasIdentifier></datacite:hasIdentifier>
    <+<datacite:hasIdentifier></datacite:hasIdentifier>
    <-<datacite:hasIdentifier>
      <-<datacite:Identifier>
        <datacite:usesIdentifierScheme rdf:resource=
        <litre:hasLiteralValue>105432011X</litre:has
        </datacite:Identifier>
      </datacite:hasIdentifier>
    <dblp:primaryFullCreatorName>Silvio Peroni</dblp:
  <-<dblp:primaryAffiliation>
    University of Bologna, Department of Classical Ph
    </dblp:primaryAffiliation>
    <dblp:otherAffiliation>OpenCitations.net</dblp:ot
    <dblp:primaryHomepage rdf:resource="http://ww
    <dblp:otherHomepage rdf:resource="https://www.
    <dblp:webpage rdf:resource="https://scholar.google
```

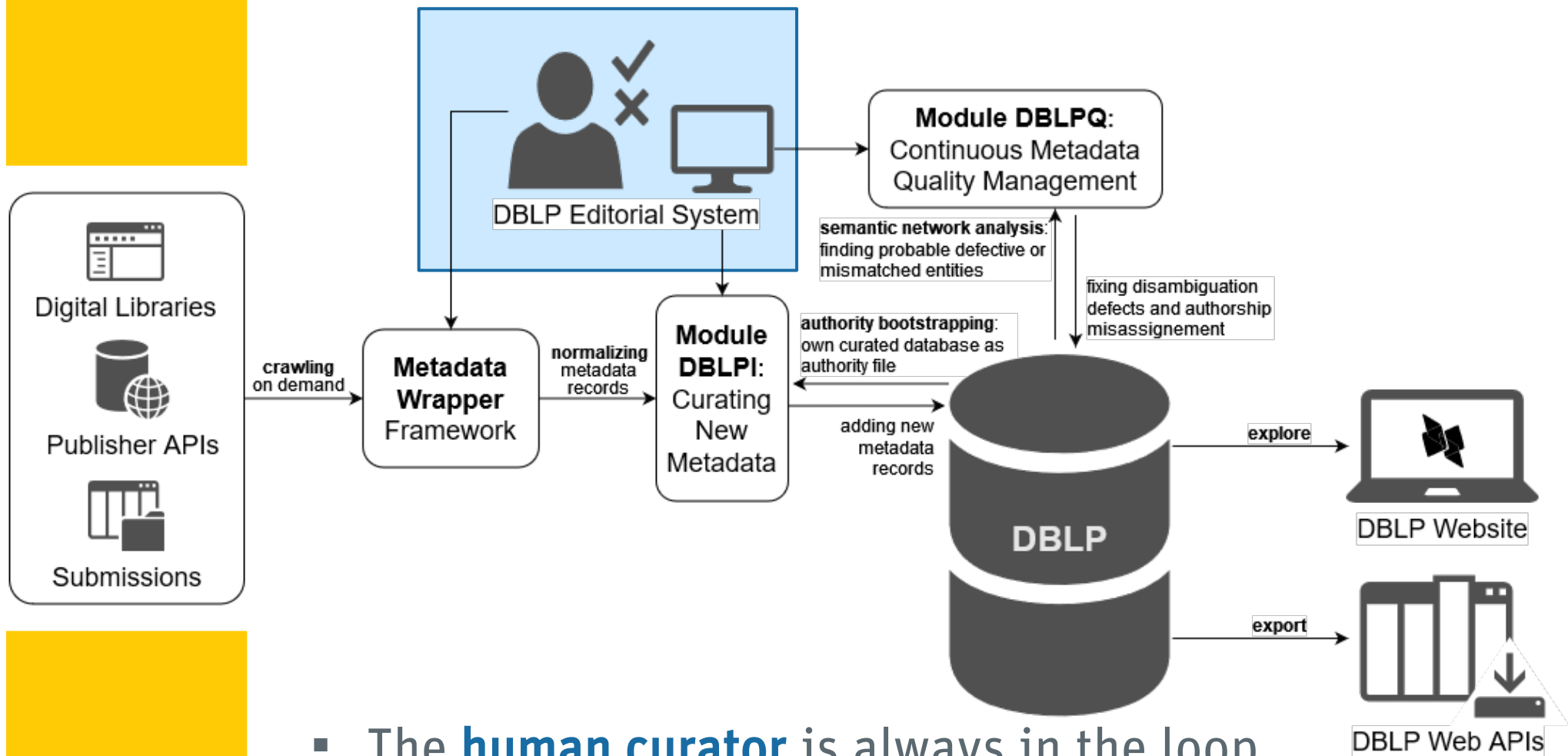
Open data APIs and full dump download of all of dblp's data for unrestricted reuse under **CCO license**

## Index of /xml/release

	<u>Name</u>	<u>Last modified</u>	<u>Size</u>
	<a href="#">Parent Directory</a>		-
	<a href="#">dblp-2022-08-01.xml.gz.md5</a>	2022-08-01 21:56	57
	<a href="#">dblp-2022-08-01.xml.gz</a>	2022-08-01 21:56	686M
	<a href="#">dblp-2022-07-01.xml.gz.md5</a>	2022-07-01 20:11	57
	<a href="#">dblp-2022-07-01.xml.gz</a>	2022-07-01 20:11	681M
	<a href="#">dblp-2022-06-01.xml.gz.md5</a>	2022-06-01 22:21	57
	<a href="#">dblp-2022-06-01.xml.gz</a>	2022-06-01 22:21	675M
	<a href="#">dblp-2022-05-02.xml.gz.md5</a>	2022-05-02 22:12	57
	<a href="#">dblp-2022-05-02.xml.gz</a>	2022-05-02 22:12	670M



# The dblp approach



- The **human curator** is always in the loop
- Algorithms provide **suggestions, not solutions**
- dblp is a **living dataset**



# Metadata wrappers for publisher data

- working with **200+ data sources**
- normalizing into internal XML format



## Digital Libraries

- largest number of wrappers
- crawlers parsing HTML websites
- esp. harvesting community-run venues



## Publisher APIs

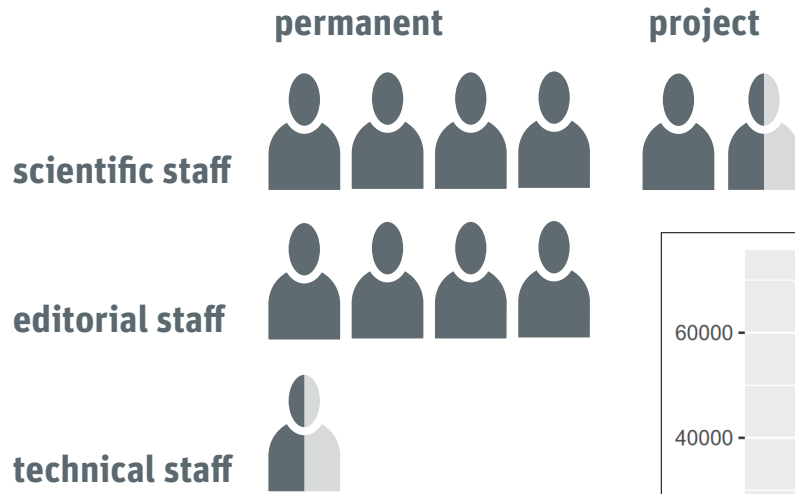
- adopted to publishers
- generic interfaces like OJS, OAI-PMH
- generic Crossref wrapper



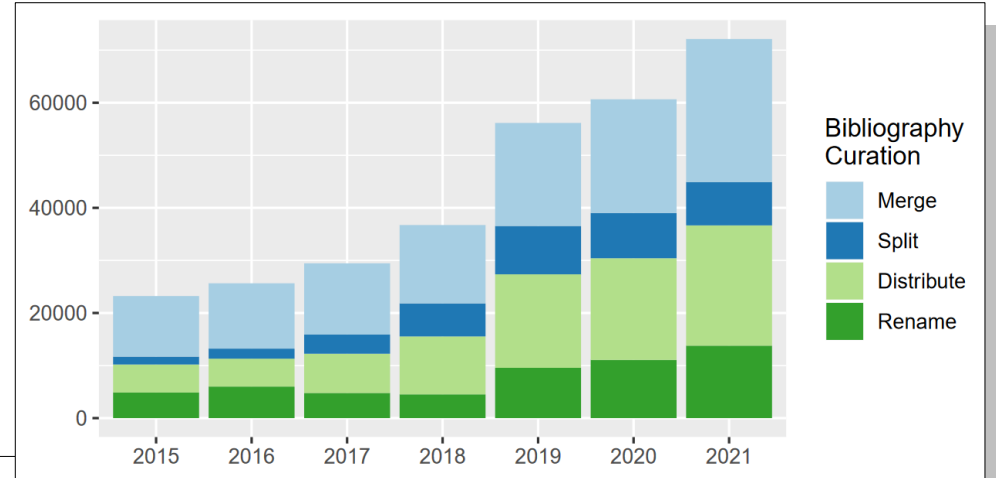
## Submissions

- data partnership with a few big publishers
- mail, FTP push/pull
- individual submissions from the community

# Data inclusion and curation efficiency

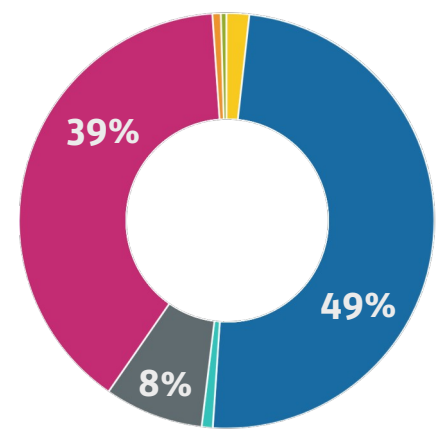
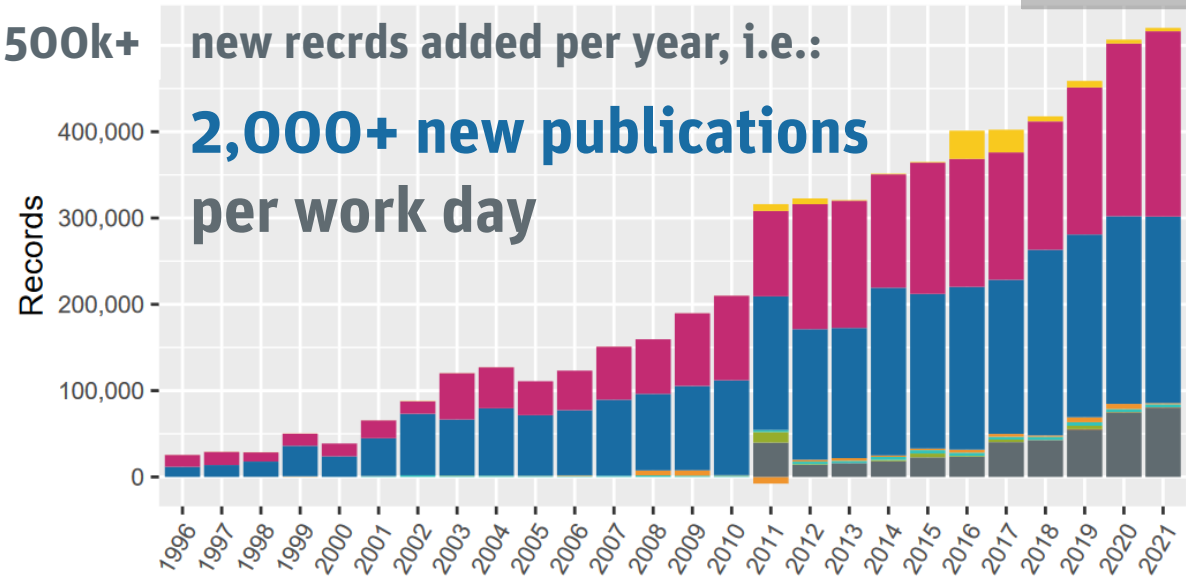


**70k+ error cases handled per year**



**500k+** new recrds added per year, i.e.:

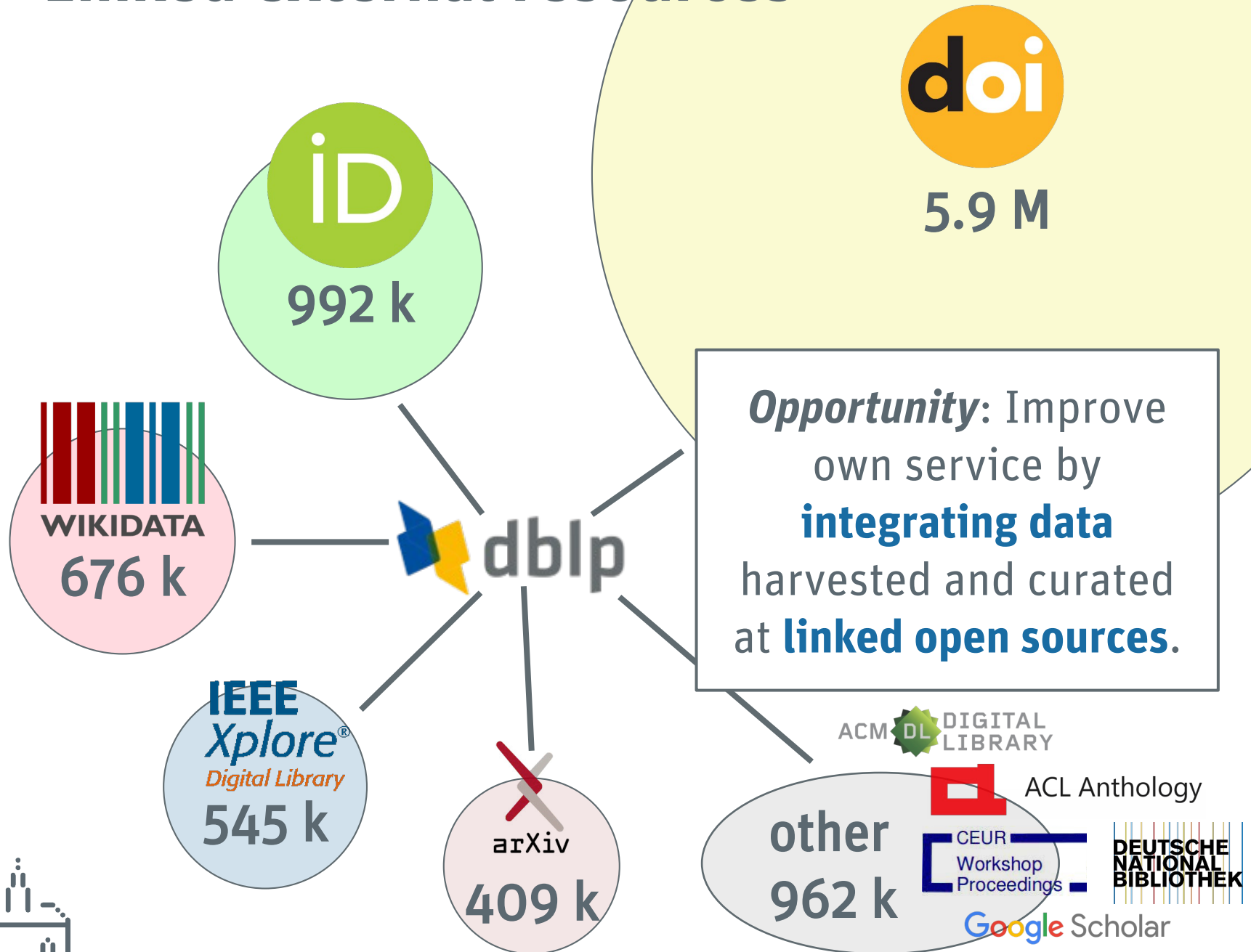
**2,000+ new publications per work day**



Marcel R. Ackermann



# Linked external resources



# Live integration of OpenAlex data



## OpenAlex API lookup:

<https://api.openalex.org/works/doi:10.4018/JSWIS.2009081901>

```
{
  id: https://openalex.org/W2015191210,
  doi: https://doi.org/10.4018/jswis.2009081901,
  title: "Linked Data - The Story So Far",
  display_name: "Linked Data - The Story So Far",
  publication_year: 2009,
  publication_date: "2009-07-01",
  ids: {
    openalex: https://openalex.org/W2015191210,
    doi: https://doi.org/10.4018/jswis.2009081901,
    mag: "2015191210"
  },
  language: "en",
  primary_location: {
    is_oa: false,
    landing_page_url: https://doi.org/10.4018/jswis.2009081901,
    pdf_url: null,
    source: {
      id: https://openalex.org/S181240966,
      display_name: "International journal on semantic web  
International journal on Semantic Web and information",
      issn_l: "1552-6283",
      issn: [
        "1552-6283",
        "1552-6291"
      ]
    }
  }
}
```

[\[-\] Details and statistics](#)

**DOI:** 10.4018/JSWIS.2009081901

**access:** closed

**type:** Journal Article

**metadata version:** 2018-11-30

load more data from openalex.org

### contained concepts

Computer science Linked data Exploit

Context (archaeology) Data science

World Wide Web Publishing

Data publishing Web standards

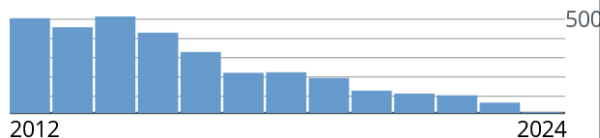
Set (abstract data type) Open data

Best practice Semantic Web

The Internet

source: [openalex.org](https://openalex.org)

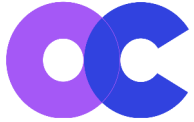
### recent citations by year



source: [openalex.org](https://openalex.org)



# Live integration of OpenCitation data



OpenCitations API lookup:  
[https://opencitations.net/index/api/v1/citations/10.1007/978-3-540-76298-0\\_52](https://opencitations.net/index/api/v1/citations/10.1007/978-3-540-76298-0_52)

[+] "DBpedia: A Nucleus for a Web of Open Data."   
 [-]   
 Sören Auer et al. (2007)

> Home

■ Sören Auer , Christian Bizer , Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, Zachary G   
 **DBpedia: A Nucleus for a Web of Open Data.** ISWC/ASWC 2007: 722-735

[+] Cited by

load citations from opencitations.net and semanticscholar.org

■ Rui Mao, Kai He, Xulang Zhang, Guanyi Chen, Jinjie Ni, Zonglin Yang, Erik Cambria :   
 **A survey on semantic processing techniques.** Inf. Fusion 101: 101988 (2024)

■ Fakhare Alam , Hamed Babaei Giglou , Khalid Mahmood Malik :   
 **Automated clinical knowledge graph generation framework for evidence based medicine.** Inf. Syst. 120964 (2023)

■ Yongsik Kwon, Seonho Lee, Yehyun Nam, Joong Chae Na, Kunsoo Park, Sang K. Cha, Bongki Moon :   
 **DB+-tree: A new variant of B+-tree for main-memory database systems.** Inf. Syst. 119: 101988 (2024)

Lan, Y., He, S., Liu, K., & Zhao, J. (2023).   
 **Knowledge Reasoning via Jointly Modeling Knowledge Graphs and Soft Rules.** Applied Science 10.3390/app131910660 [\[edit\]](#)

HUANG, X., LIU, Z., ZHANG, M., 桂韬, 韩先培, 户保田, 黄民烈, 黄萱菁, 刘康, 刘挺, 刘知远, 秦兵, 邱锡鹏, 万小 张民, ... 赵妍妍. (2023). Towards a comprehensive understanding of the impact of large language processing: challenges, opportunities and future directions<sup>†</sup> SCIENTIA SINICA Informationis, ssi-2023-0113 [\[edit\]](#)

Zhou, X., Zhang, S., Agarwal, M., Akroyd, J., Mosbach, S., & Kraft, M. (2023).   
 **Marie and BERT—A Knowledge Graph Embedding Based Question Answering System for Knowledge Graphs.** Inf. Syst. 120964 (2023). doi:10.1021/acsomega.3c05114 [\[edit\]](#)

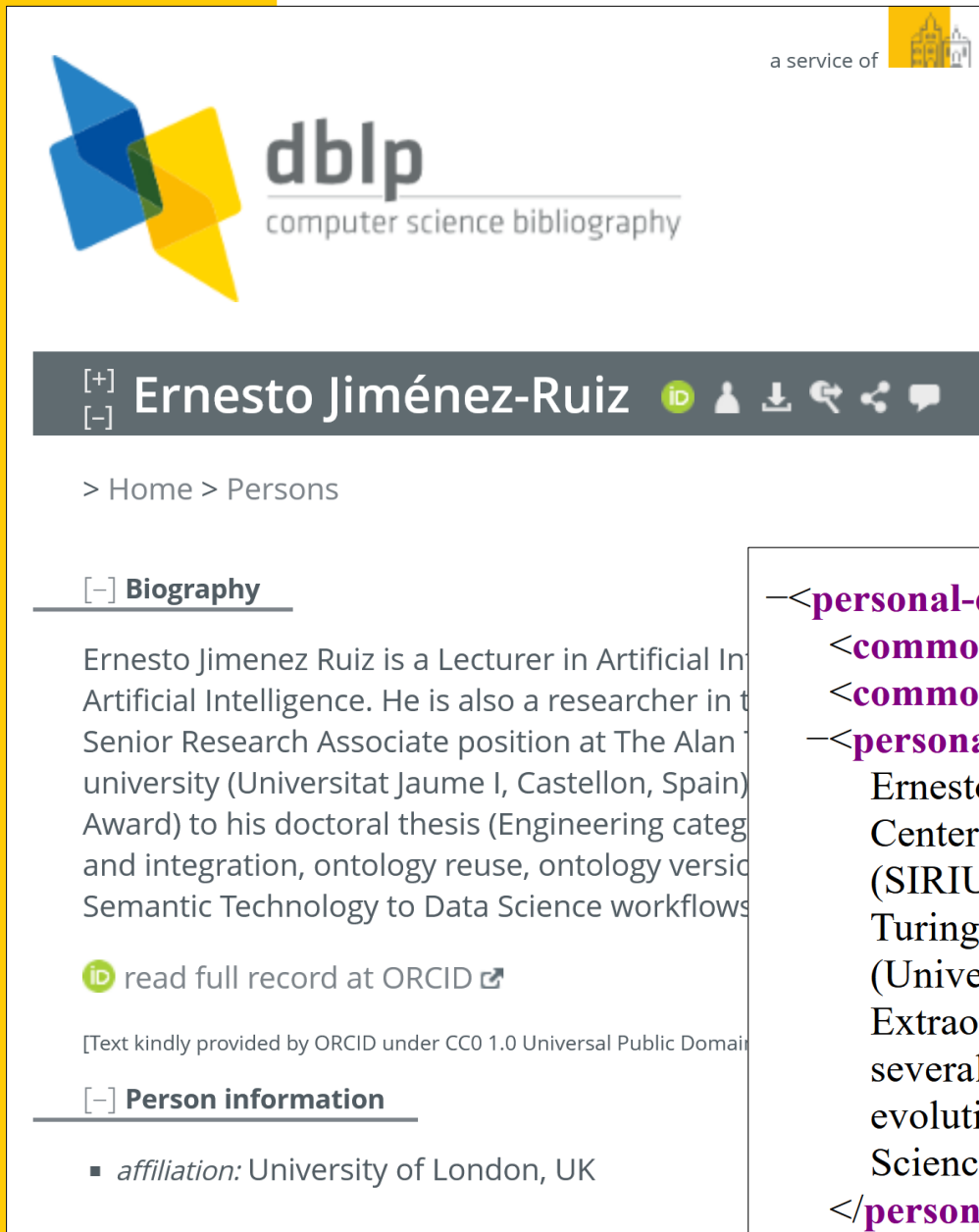
■ Xue Zhou , Bei Hui, Ilana Zeira, Hao Wu, Ling Tian:   
 **Dynamic relation learning for link prediction in knowledge hypergraphs.** Appl. Intell. 53(1): 1-15 (2023)

■ Weiwei Sun , Shuyu Guo , Shuo Zhang , Pengjie Ren , Zhumin Chen , Maarten de Rijke :   
 **Metaphorical User Simulators for Evaluating Task-oriented Dialogue Systems.** ACM Trans. Inf. Syst. 42(1): 1-25 (2022)

```
{
  "journal_sc": "no",
  "oci": "06101160324-06503267503",
  "author_sc": "no",
  "citing": "10.1109/access.2019.2904835",
  "timespan": "P12Y",
  "cited": "10.1007/978-3-540-76298-0_52",
  "creation": "2019"
},
{
  "journal_sc": "no",
  "oci": "06101305463-06503267503",
  "author_sc": "no",
  "citing": "10.4018/ijssci.301567",
  "timespan": "P15Y",
  "cited": "10.1007/978-3-540-76298-0_52",
  "creation": "2022"
},
{
  "journal_sc": "no",
  "oci": "06101482253-06503267503",
  "author_sc": "no",
  "citing": "10.3333/ifc.151025"
```



# Live integration of ORCID biographies



The screenshot shows the dblp profile for Ernesto Jiménez-Ruiz. At the top left is the dblp logo (computer science bibliography) and a small 'a service of' logo. The profile name 'Ernesto Jiménez-Ruiz' is displayed with an ORCID icon and social media icons. Below the name is a breadcrumb '> Home > Persons'. A section titled 'Biography' contains a paragraph of text about his role as a Lecturer in Artificial Intelligence and Senior Research Associate at the University of Jaume I. Below the biography is a link 'read full record at ORCID'. A section titled 'Person information' lists his affiliation as 'University of London, UK'.

ORCID

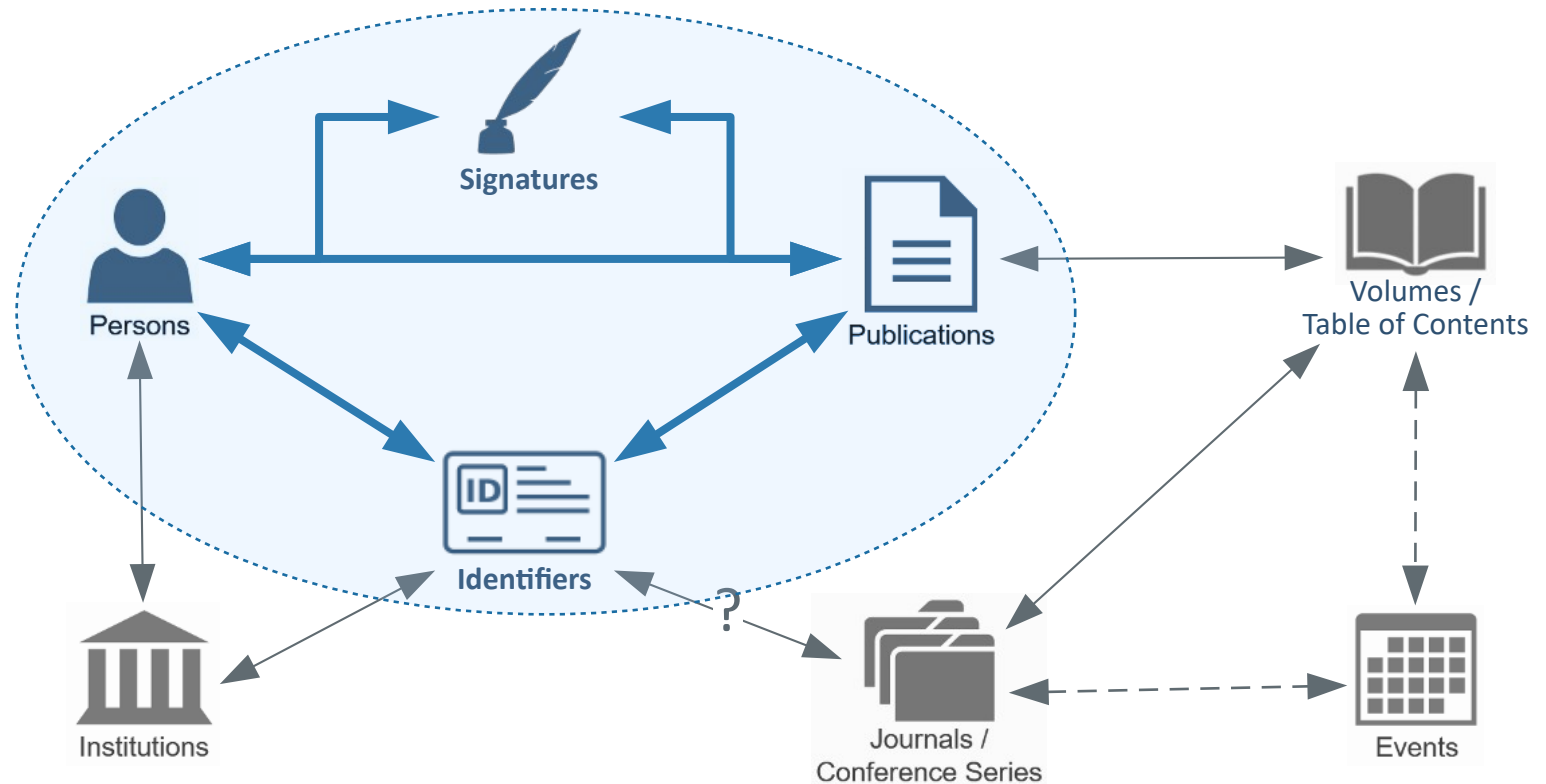
ORCID API lookup:

<https://pub.orcid.org/v3.0/0000-0002-9083-4599/biography>

```
<personal-details:biography visibility="public" path="/0000-0002-9083-4599/biography">
  <common:created-date>2016-04-15T23:17:35.454Z</common:created-date>
  <common:last-modified-date>2020-01-10T17:22:24.699Z</common:last-modified-date>
  <personal-details:content>
    Ernesto Jimenez Ruiz is a Lecturer in Artificial Intelligence at City, University of London, UK and a Senior Research Associate at the Norwegian Research Centers for Machine Learning and Artificial Intelligence. He is also a Senior Research Associate at the Norwegian Research Centers (SIRIUS) at the University of Oslo, Norway. He previously held a Senior Research Associate position at the Alan Turing Institute in London (UK) and a Research Assistant position at the University of Jaume I, Castellon, Spain) awarded a “Premio extraordinario (Extraordinary Doctoral Award) to his doctoral thesis (Engineering category) in several areas, including bio-medical information processing and integration, ontology evolution, ontology alignment. His current research interests focus on Knowledge Representation, Semantic Technology to Data Science workflows and the combination of Knowledge Representation
  </personal-details:content>
</personal-details:biography>
```



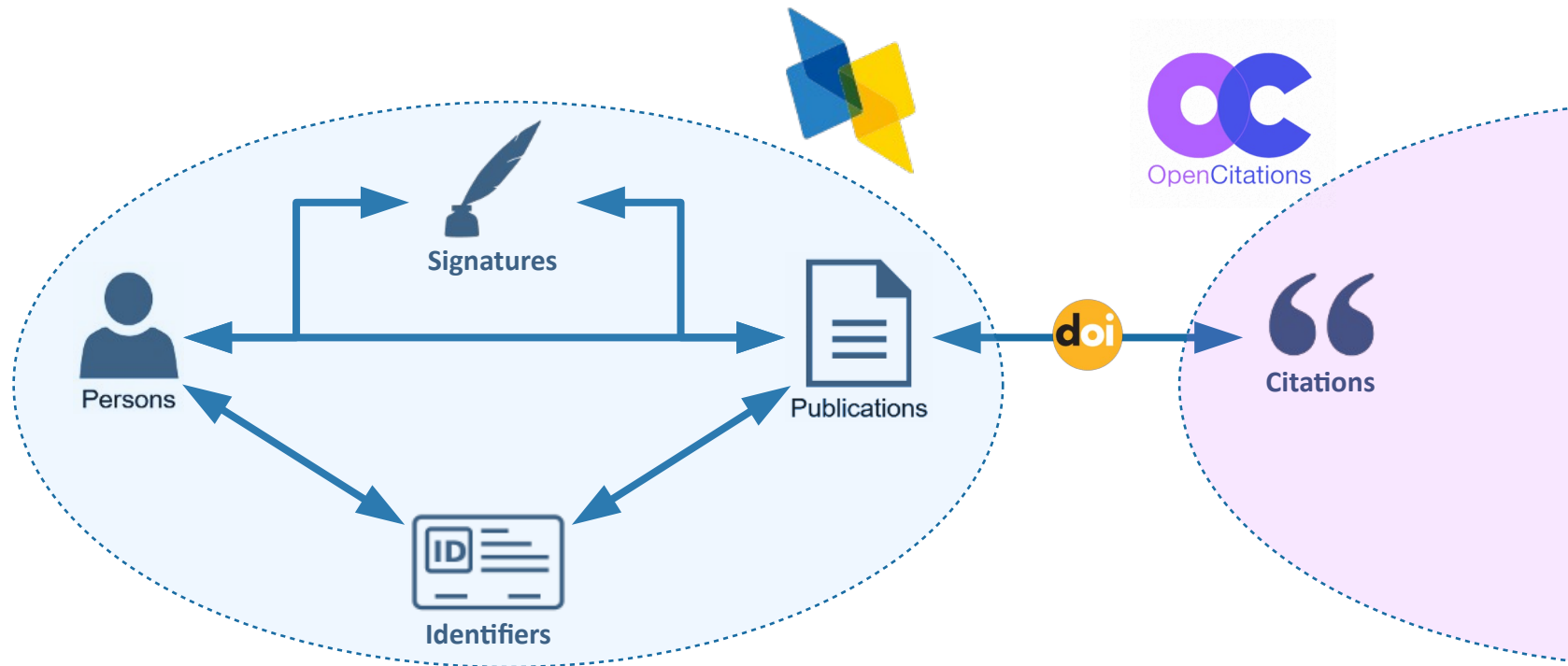
# dblp as a knowledge graph



- **3.4 million** Persons
- **7.1 million** Publications
- **23 million** Signatures
- **18 million** Identifiers
- **388 million** triples
- **9.5 million** linked external URIs



# Joint dataset: dblp + OpenCitations



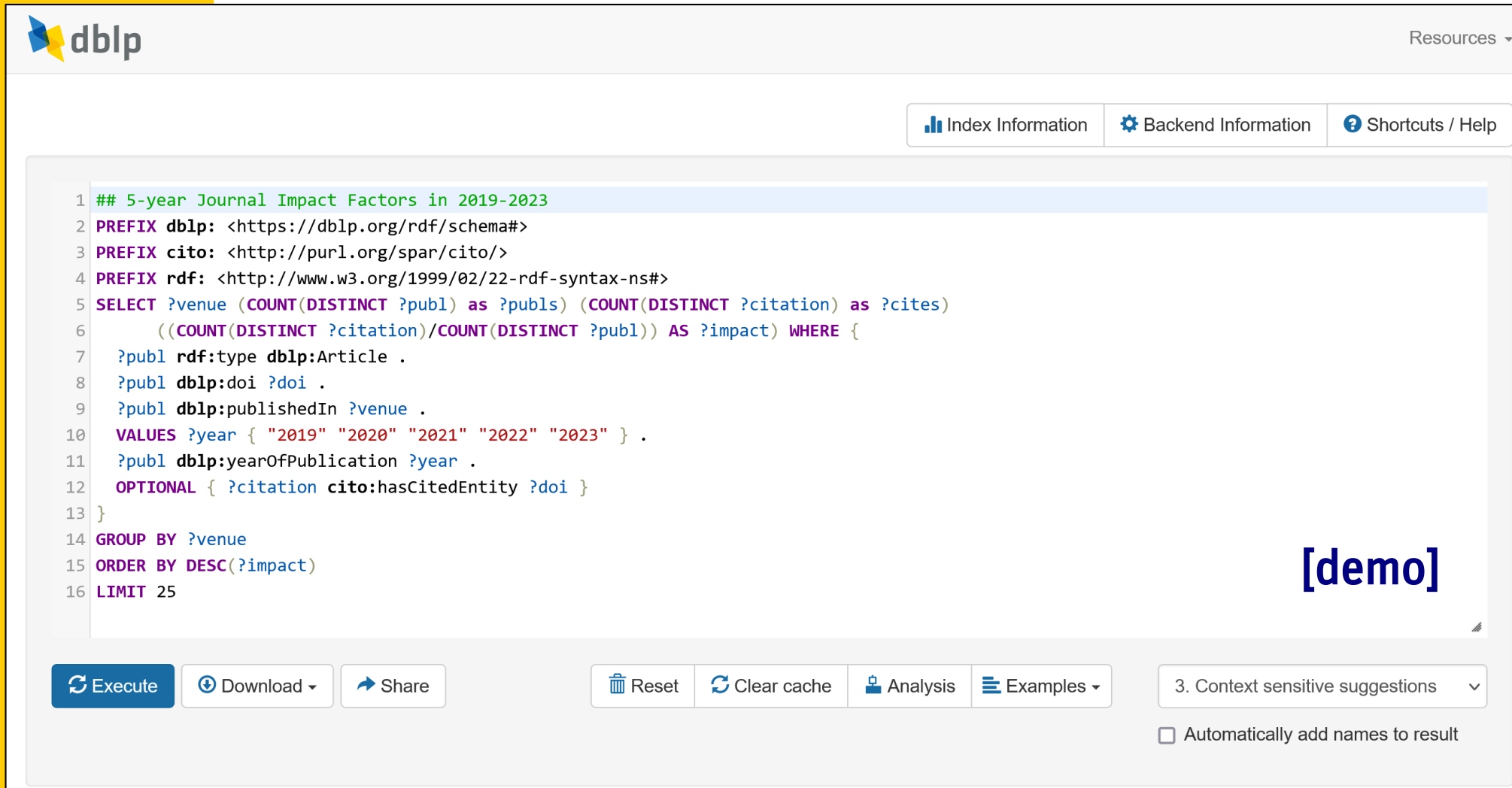
Adding all COCI+DOCI citations with DOI in dblp

- **122 million** Citations
- **1.15 billion** triples in joint dataset



# Public SPARQL endpoint

- <https://sparql.dblp.org> (almost public beta)








The screenshot shows the dblp SPARQL endpoint interface. At the top left is the dblp logo, and at the top right is a 'Resources' dropdown menu. Below the logo are three navigation buttons: 'Index Information', 'Backend Information', and 'Shortcuts / Help'. The main area contains a SPARQL query editor with a light blue background. The query is as follows:

```
1 ## 5-year Journal Impact Factors in 2019-2023
2 PREFIX dblp: <https://dblp.org/rdf/schema#>
3 PREFIX cito: <http://purl.org/spar/cito/>
4 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
5 SELECT ?venue (COUNT(DISTINCT ?publ) as ?publs) (COUNT(DISTINCT ?citation) as ?cites)
6     ((COUNT(DISTINCT ?citation)/COUNT(DISTINCT ?publ)) AS ?impact) WHERE {
7   ?publ rdf:type dblp:Article .
8   ?publ dblp:doi ?doi .
9   ?publ dblp:publishedIn ?venue .
10  VALUES ?year { "2019" "2020" "2021" "2022" "2023" } .
11  ?publ dblp:yearOfPublication ?year .
12  OPTIONAL { ?citation cito:hasCitedEntity ?doi }
13 }
14 GROUP BY ?venue
15 ORDER BY DESC(?impact)
16 LIMIT 25
```

At the bottom right of the query editor, there is a large blue button labeled '[demo]'. Below the query editor is a control bar with several buttons: 'Execute', 'Download', 'Share', 'Reset', 'Clear cache', 'Analysis', and 'Examples'. To the right of these buttons is a dropdown menu showing '3. Context sensitive suggestions' and a checkbox labeled 'Automatically add names to result'.




# dblp website integration

[+] Rüdiger Reischuk     

[-] K. Rüdiger Reischuk

> Home > Persons

[-] SPARQL queries 

---





**run query for this person**

- highly cited coauthors
- co-cited other authors
- non-coauthors with many
- highly cited publications
- co-cited publications


*or build your own?*

[demo]



[+] Symposium on the Theory of    

> Home > Conferences and Wo

[-] SPARQL queries 

---

**run query for this venue**


- frequent authors
- venues with community over
- co-cited venues
- venues citing this venue
- highly cited publications

*or build your own?*

[demo]

[+] "Fully homomorphic encr  
[-] Craig Gentry (2009)

> Home

[-] SPARQL queries 

---

**run query for this record**

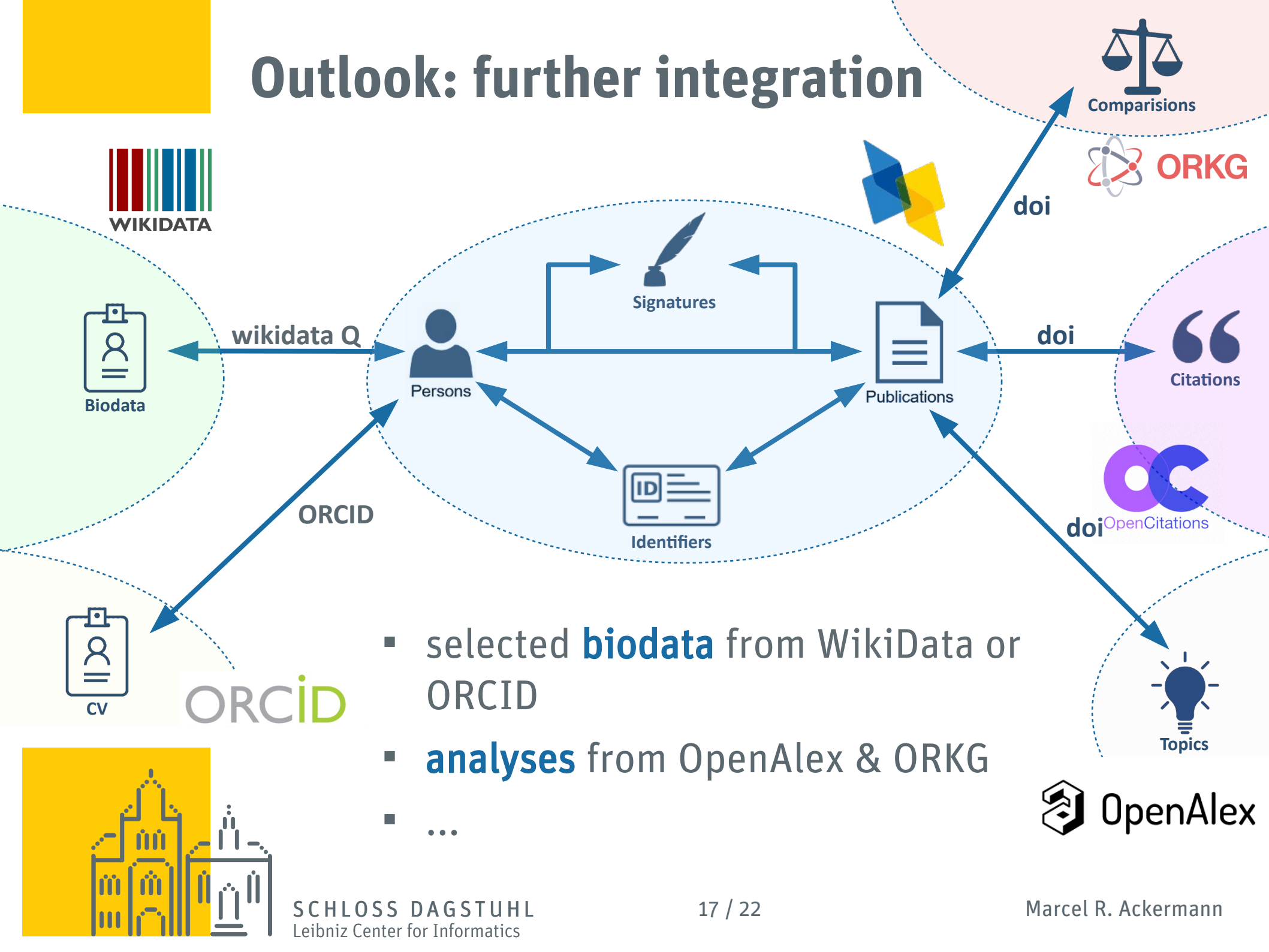
- referenced publications
- cited by publications
- co-cited publications
- citing venues
- co-cited venues

*or build your own?*

[demo]



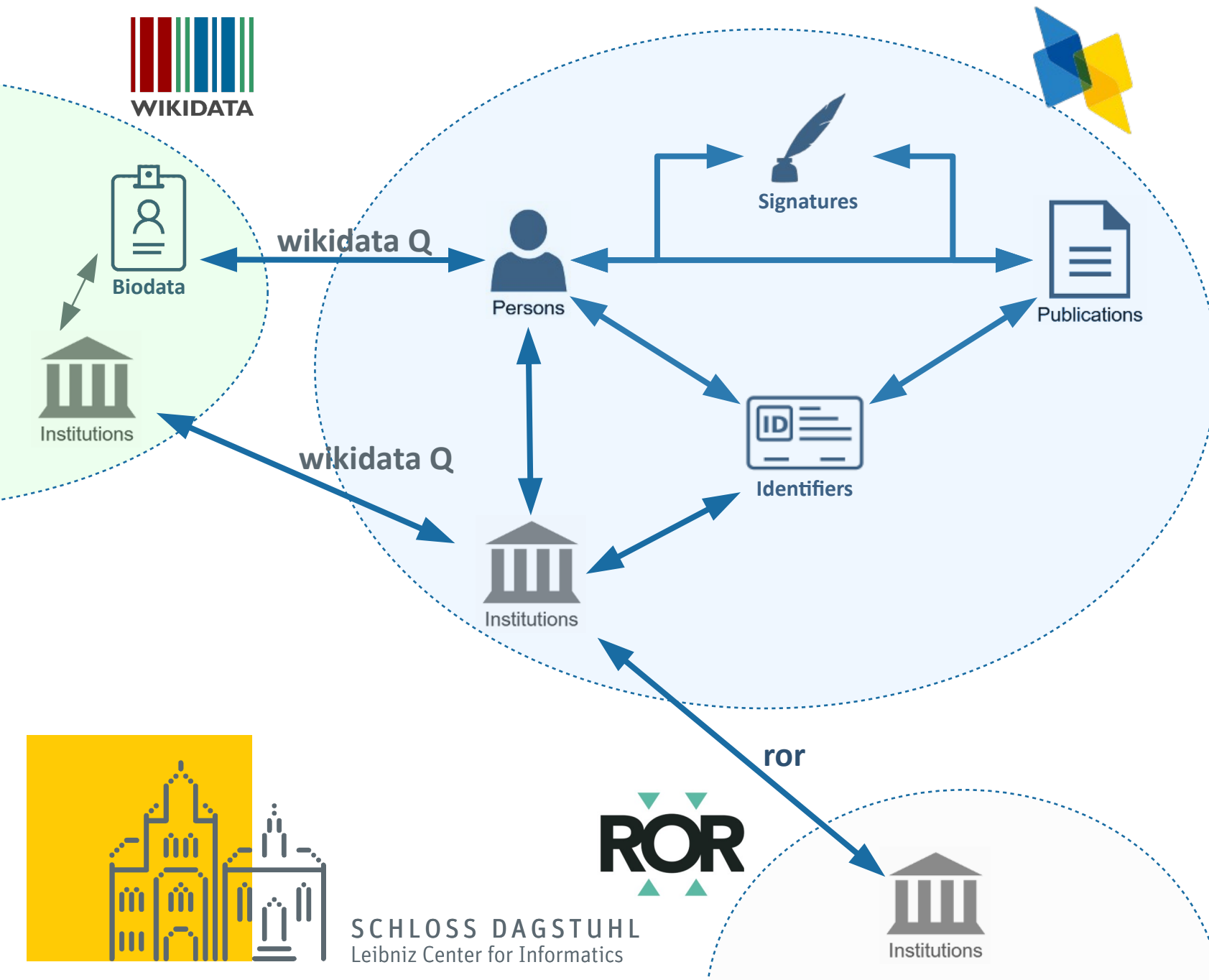
# Outlook: further integration







- selected **biodata** from WikiData or ORCID
- **analyses** from OpenAlex & ORKG
- ...




# Outlook: expanding the dblp schema



# Affiliations and institution metadata


[+] [-] Massachusetts Institute of Technology      
MIT

> Home > Institutions > US Trier 2




[+] Information 

**Massachusetts Institute of Technology (MIT)** is a private research university in Cambridge, Massachusetts. The institute is a land-grant, sea-grant, and space-grant university, with an urban campus that extends more than a mile (1.6 km) alongside the Charles River. The institute also encompasses a number of major off-campus facilities such as the MIT Lincoln Laboratory, the Bates Center, and the Haystack Observatory, as well as affiliated laboratories such as the

[+] Map




Leaflet | Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap










[+] [-] Nicolas Bourbaki   

a) [-] Person information

- *affiliation*: University of Berlin, Germany
- *affiliation [computed]*: University of Paris, France
- *affiliation [computed]*: University of London, UK
- *affiliation (PhD 2005)*: University of Madrid, Spain

[+] 2020 - today 

2020

- [c6]    Christine Geeng, Nicolas Bourbaki: **EGregor: An Eldritch Privacy Mental Model for Smart Assistants.** CHI Extended Abstracts 2020: 1-9
- [c5]    Nicolas Bourbaki, Arian Akhavan Niaki, Nguyen Phong Hoang, Phillipa Gill, Amir Houmansadr: **Triplet Censors: Demystifying Great Firewall's DNS Censorship Behavior.** FOCI @ USENIX Security Symposium 2020
- [c4]    Nicolas Bourbaki: **Real or surreal: A pilot study on creative idea generation in MR vs. VR.** VR Workshops 2020: 677-678

[+] Refine list

showing all 57 records

refine by search term

refine by type ...

refine by coauthor  
Jane Doe (8)  
*more options*

refine by venue  
TPDL (5)  
*more options*




b) refine by affiliation  
University of Paris (11)  
University of Berlin (9)  
University of London (9)  
University of Madrid (7)  
Affiliation unknown (21)

**DFG**

*SmartER Affiliations*  
*Harvesting of Author Affiliation Data*

- Oct 2023 – Sep 2026
- 0.5 FTE since Nov 2023

# Affiliation data sources

- Current prototype:
  - affiliations for <20% of all author signatures
  - based on Wikidata & old MS Academic Graph
- Future sources
  - publisher metadata,  Crossref,  OpenAlex, ...
  - mining from papers
  - **persistent ID** provider: 

## Graph Threading

Erik D. Demaine  
MIT

Yael Kirkpatrick  
MIT

Rebecca Lin  
MIT

### Abstract

Inspired by artistic practices such as beadwork and himmeli, we study the problem of *threading* a single string through a set of tubes, so that pulling the string forms a desired graph. More



# To do: better entity recognition

Search the registry

MIT



30 Organizations

<https://ror.org/047m6hg73>

Management Intelligenter Technologien (Germany)

OTHER

ORGANIZATION

Corporation

<https://ror.org/04mtcj695>

University of Southern Mindanao

OTHER NAMES

ORGANIZATION

Education

<https://ror.org/02qk9nj07>

Ministry of Infrastructures and Transport

OTHER

ORGANIZATION

Government

<https://ror.org/02mb6z761>

Myanmar Institute of Theology

OTHER NAMES: Baptist Divinity School, MIT

ORGANIZATION TYPES

Education

<https://ror.org/002r67t24>

Manukau Institute of Technology

OTHER NAMES:

ORGANIZATION

Education

<https://ror.org/042nb2s44>

Massachusetts Institute of Technology

OTHER NAMES: Instituto Tecnológico de Massachusetts (es), MIT

ORGANIZATION

Education

<https://ror.org/00v140q16>

MIT University

OTHER NAMES: МИТ Универзитет (mk)

ORGANIZATION TYPES

Education

RELATIONSHIPS

View details

LOCATIONS

Yangon (GeoNames)

WEBSITE

RELATIONSHIPS

View details

View details

View



# Thanks!

<https://dblp.org>

✉ [dblp@dagstuhl.de](mailto:dblp@dagstuhl.de)

~~ @dblp\_org~~

 [@dblp@mastodon.social](https://mstdn.social/@dblp)

