ORCID for Wikidata

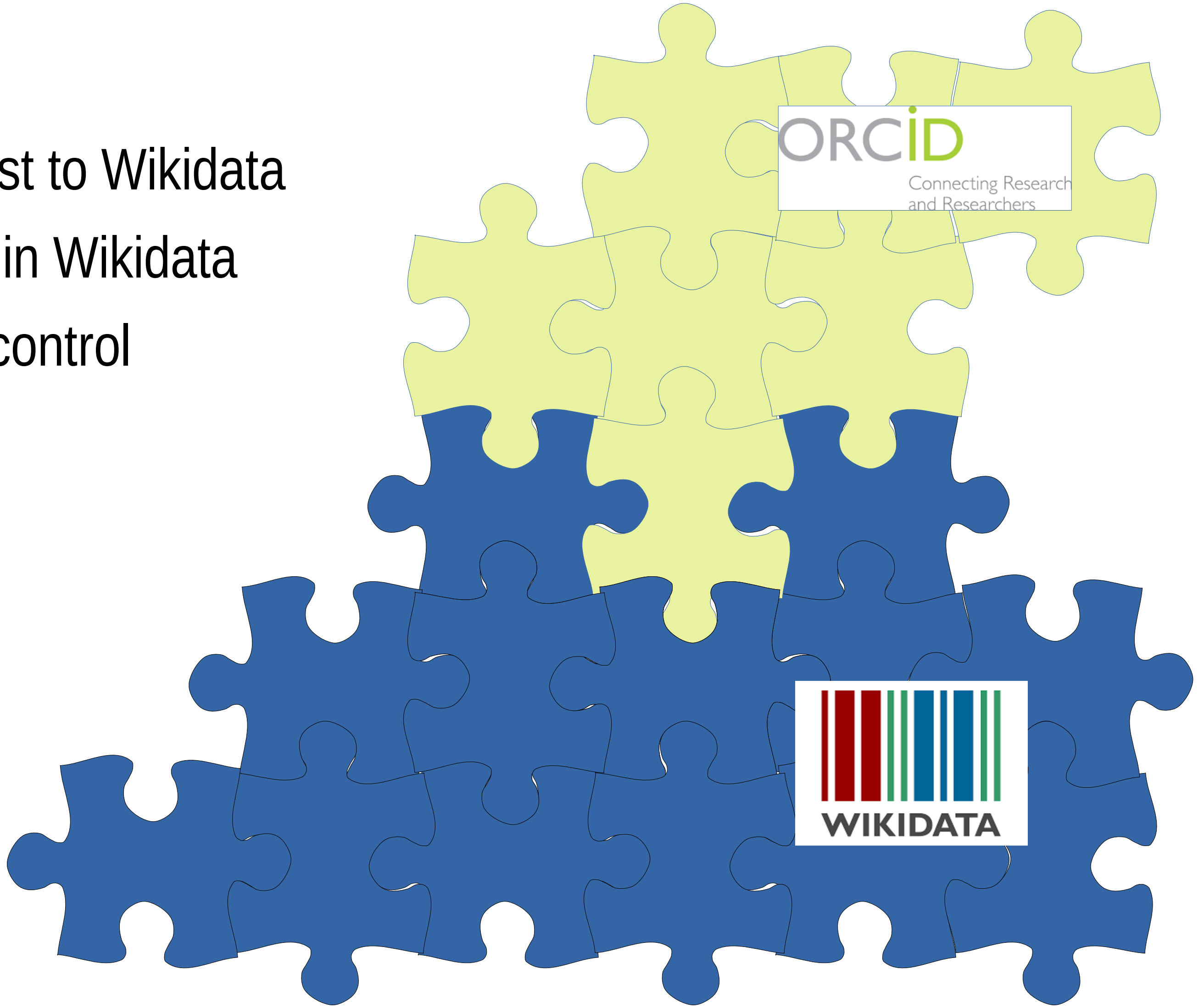
Improving bibliometric data in Wikidata

Dr. Eva Seidlmayer, SWIB2020 November 26, 2020

ZB MED Information Centre for Life Sciences

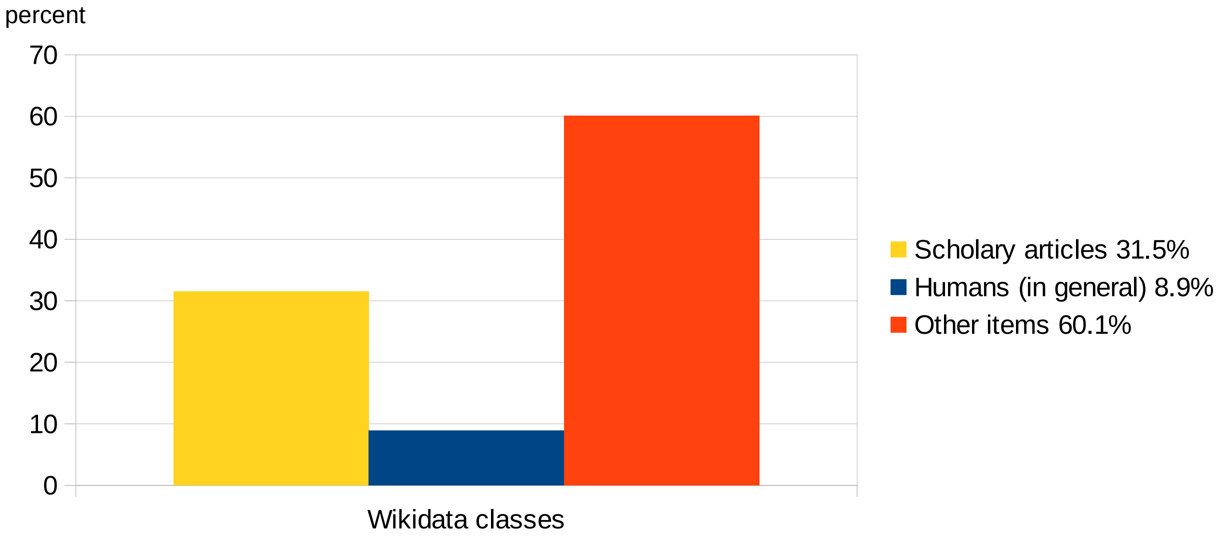
Agenda

- Introduction: Wikidata, ORCID
- Preparation of information from ORCID for ingest to Wikidata
- “OrcBot“ – a bot for densification of information in Wikidata
    - Wikidata bot: large scale upload and quality control
- Results on ORCID for Wikidata



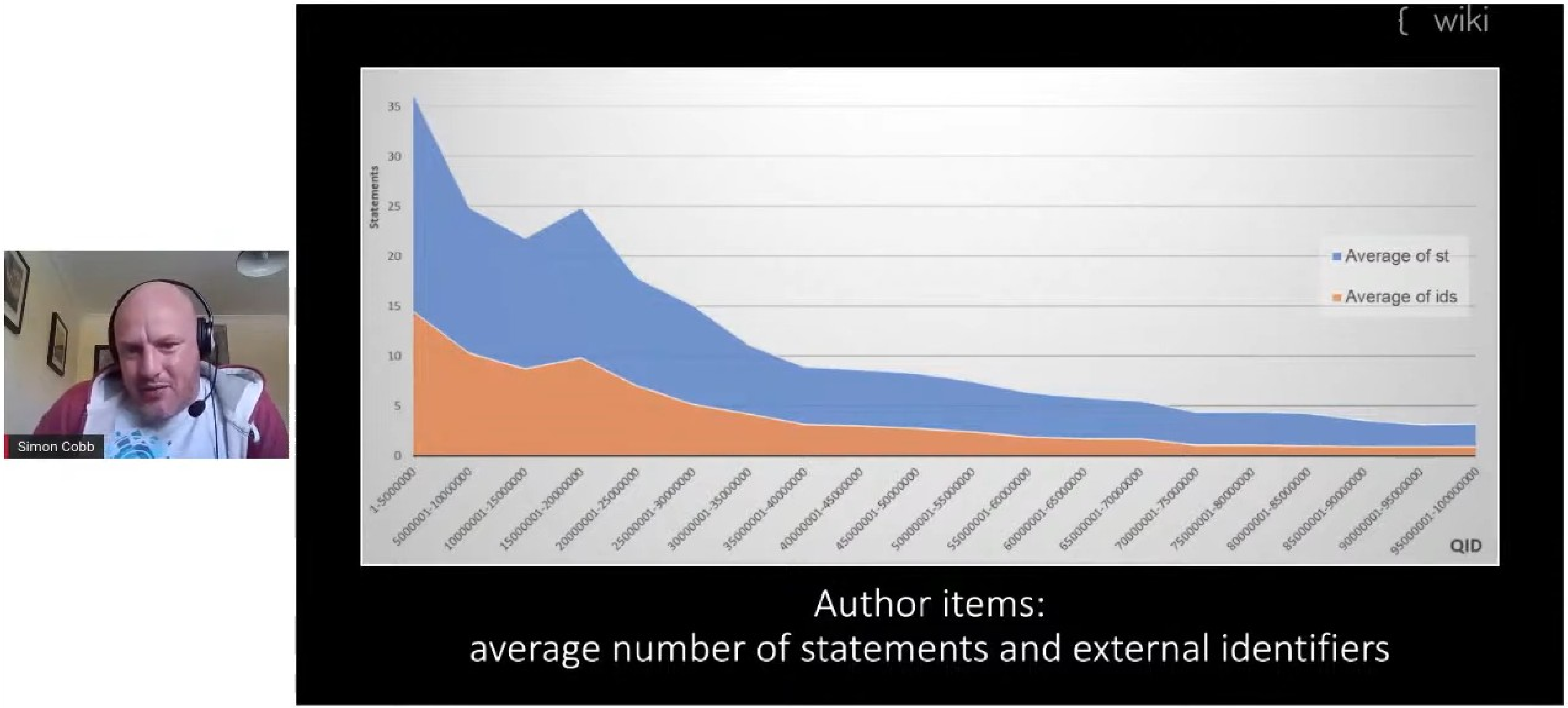
Introducing Wikidata

- Open knowledge base for semantic data
- Central storage for structured data used in Wikimedia sister projects such as: Wikipedia, Wikivoyage, Wiktionary, Wikisource und andere
- Community curated data
- Currently, 71M items (Q-ID) in Wikidata*:
  - Scholarly articles: 22.5M (31.5%)
  - Humans: 6.3M (8,9%)



*Wikidata statistics: https://www.wikidata.org/wiki/Wikidata:Statistics, status: 2020-02-16.

Introducing Wikidata



Author items:
average number of statements and external identifiers

Minute 35’-40’ in
WikiCite presentatior

Many items have
few (<10) statements!

Older author items → Recent author items

Tendency: declining richness of information

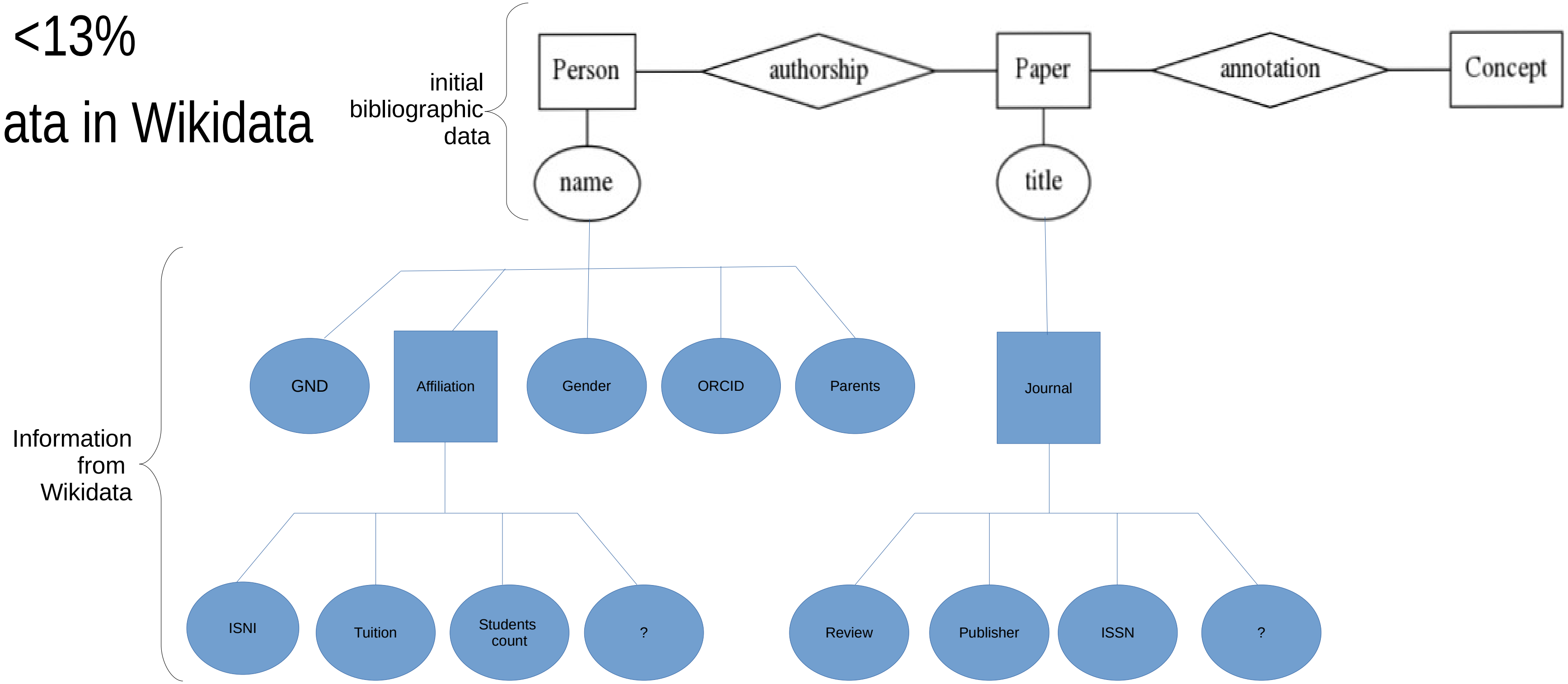
Simon Cobb, Author items, YouTube channel Wikipedia Weekly, https://youtu.be/wZUB62hp5dU, 2020-10-28.

Introducing Wikidata

Use case: “Q-Aktiv“ project by Kiel University, ZBW Kiel and ZB MED Cologne

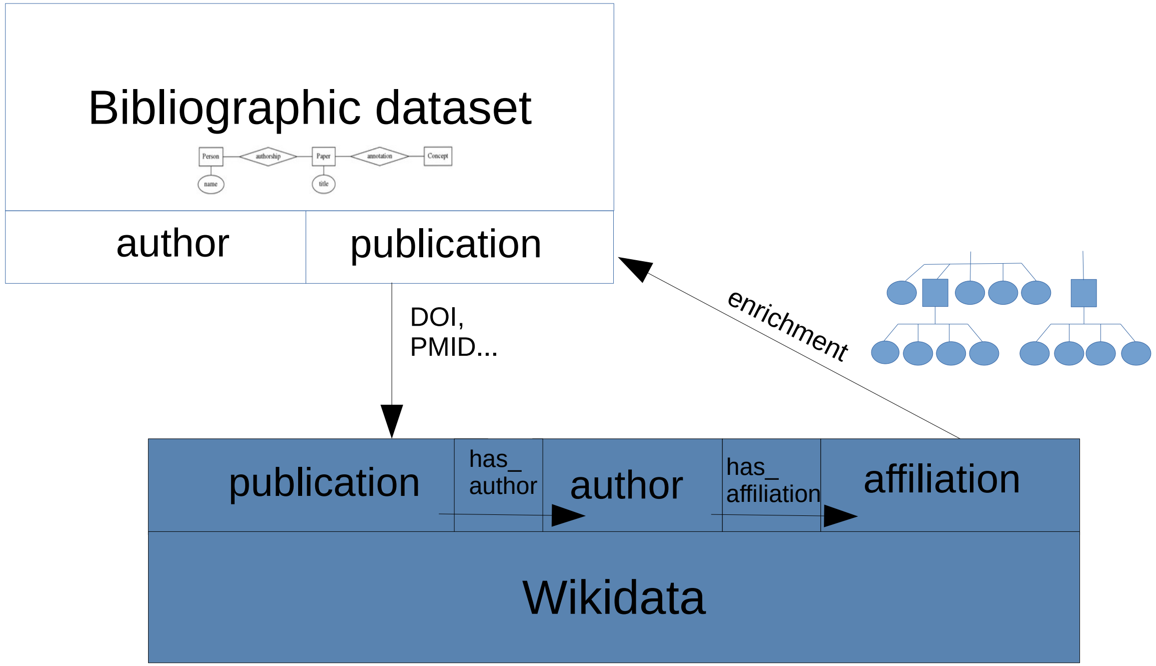
- Enrichment of bibliographic data set using Wikidata API
- In general low coverage: <13%

→ decision to improve data in Wikidata



Introducing Wikidata

- Reasons for low coverage:
  - Unstable performance of Wikidata API when requesting huge amounts of queries
  - Missing publication items (Q-ID) in Wikidata
  - Missing author items (Q-ID) in Wikidata
  - **Missing relations (has_author = P50) between author items and publication items**



→Publication/author pairs would improve the existing data in Wikidata

Introducing ORCID

- Persistent digital identifier for researchers and contributors
- Information on scientific biographies (publications, study, employer, funding...)
- Provision of information by researchers themselves: high quality of data
- CC0 license allows reuse of public data
- Currently*:
  - 9.8M researchers carry an ORCID iD
  - 62.8M registered own works (publications)

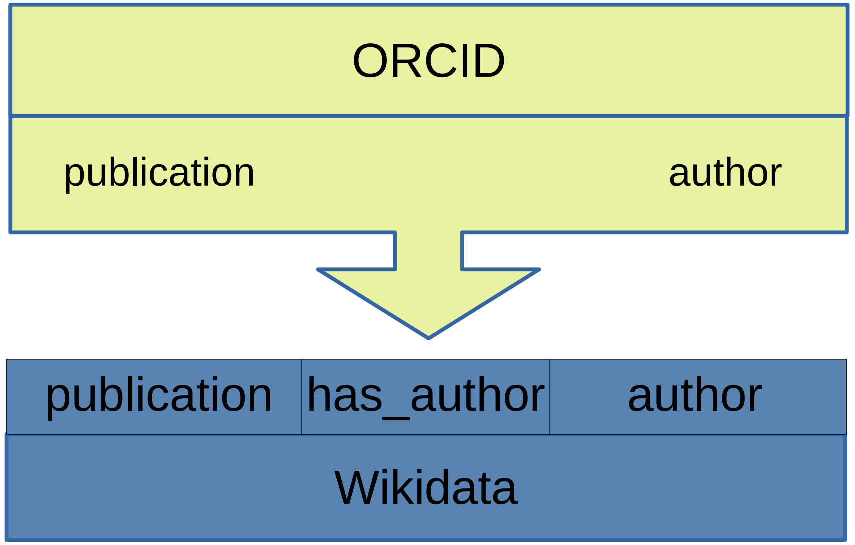
    → publication/author pairs

*ORCID statistics: https://orcid.org/statistics, status: 2020-10-24.

Introducing ORCID

- ORCID data can be used to
  - Create publication items
  - Create author items
  - **Relate publication items to author items (P50)**

→ In order not to flood Wikidata with less important items we focus on the matching of existing items



Preparation of ORCID data

- Publication dataset:
  - Harvesting of publications IDs from ORCID:

    PMID, PMC, DOI, EID, DNB, (WOS)
    - Check if publications are already registered in Wikidata:

      + publication Q-ID
      - Check if authors of publications are already registered in Wikidata:

        + all author’s Q-IDs (no author name strings (P2093))

Information from ORCID database						Information from Wikidata	
ORCID	pmid	pmc	doi	eid	dnb	publication Q ID	all_authors_Q ID
0000-0003-2760-1191	27899851.0	PMC5126442				Q37448620	Q83234445
0000-0001-7526-5191	21033872.0		10.1063/1.3475729			Q82227134	Q57037275, Q82227128
0000-0002-7871-5191	28338018.0	10.1038/srep45089	10.1038/srep45089			Q42129196	Q57212955, Q59701610, Q87372245

Preparation of ORCID data

- Author dataset
- Harvesting authors from ORCID:

  ORCID-ID, first name, last name, affiliation, affiliation ID, Affiliation ID source (e.g. Ringold, GRID), start date

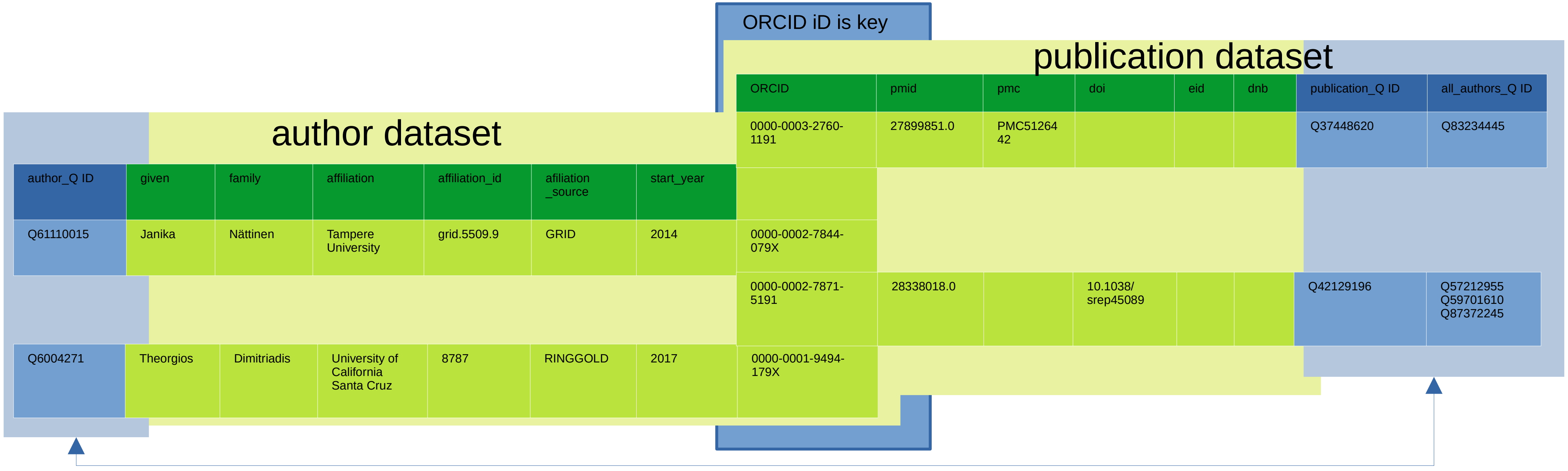
  - Check if authors are already registered to Wikidata

    + author Q-ID

Information from ORCID database							Info from Wikidata
ORCID	given	family	affiliation	affiliation_id	affiliation_source	start_year	author Q-ID
0000-0002-7844-079X	Janika	Nättinen	Tampere University	grid.5509.9	GRID	2014	Q61110015
0000-0002-0171-879X	Barbara	van Asch	Stellenbosch University	26697	RINGGOLD	2015	Q54452584
0000-0001-9494-179X	Georgios	Dimitriadis	University of California Santa Cruz	8787	RINGGOLD	2017	Q60042671

Ingest to Wikidata: OrcBot

- OrcBot combines both prepared data sets: ORCID iD is key
- Check if author is already registered as author to the publication



publication dataset

ORCID	pmid	pmc	doi	eid	dnb	publication_Q ID	all_authors_Q ID
0000-0003-2760-1191	27899851.0	PMC5126442				Q37448620	Q83234445
0000-0002-7844-079X							
0000-0002-7871-5191	28338018.0		10.1038/srep45089			Q42129196	Q57212955 Q59701610 Q87372245
0000-0001-9494-179X							

author dataset

author_Q ID	given	family	affiliation	affiliation_id	afiliation_source	start_year	ORCID
Q61110015	Janika	Nättinen	Tampere University	grid.5509.9	GRID	2014	0000-0002-7844-079X
							0000-0002-7871-5191
Q6004271	Theorgios	Dimitriadis	University of California Santa Cruz	8787	RINGGOLD	2017	0000-0001-9494-179X

OrcBot

Ingest to Wikidata: OrcBot

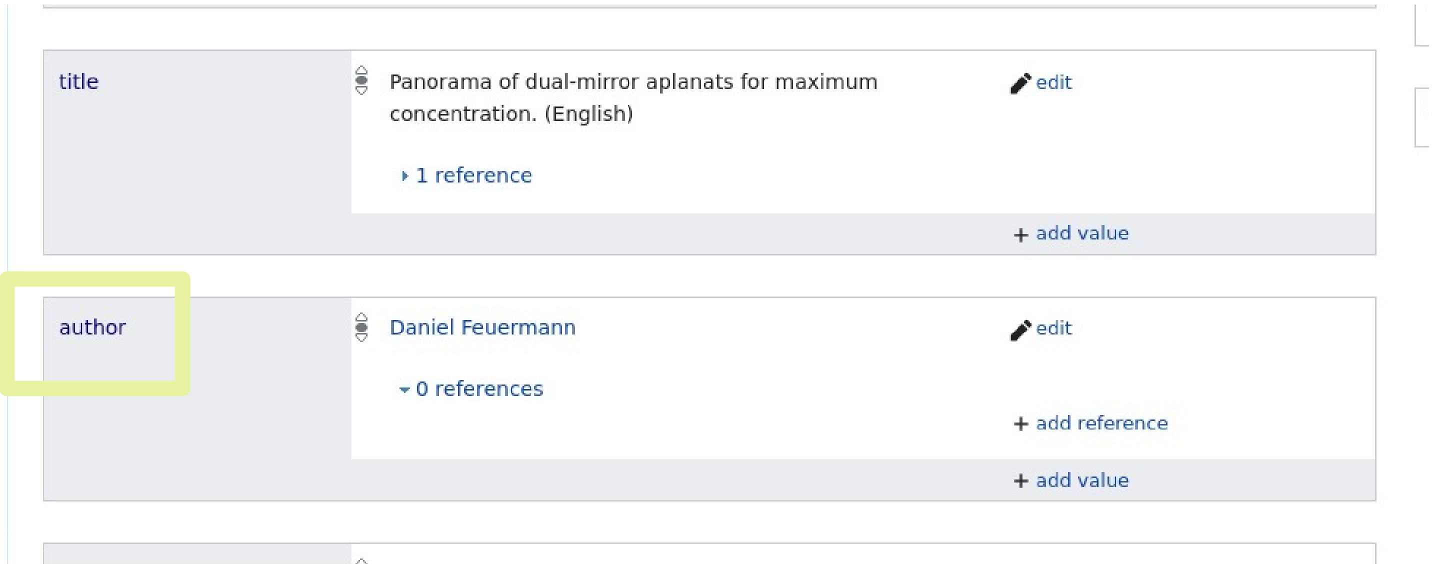
- If the author is not yet part of the list of all authors of the article in Wikidata OrcBot creates a JSON template containing author information

```
{‘id’: article-QID, # article QID
        ’claims’: {
                ‘P50’: { # has author
                        ‘value’: author-QID,  # author QID
                ‘qualifier’: [{‘P1932’:       # has author string
                        ( ‘given_name’, ‘family_name’ }]
                }}
}
```

- Upload of JSON using Wikidata CLI tool

```
wb edit-entity tmp.json
```

has_author
P50



title | Panorama of dual-mirror aplanats for maximum concentration. (English) | edit
1 reference
+ add value

author | Daniel Feuermann | edit
0 references
+ add reference
+ add value

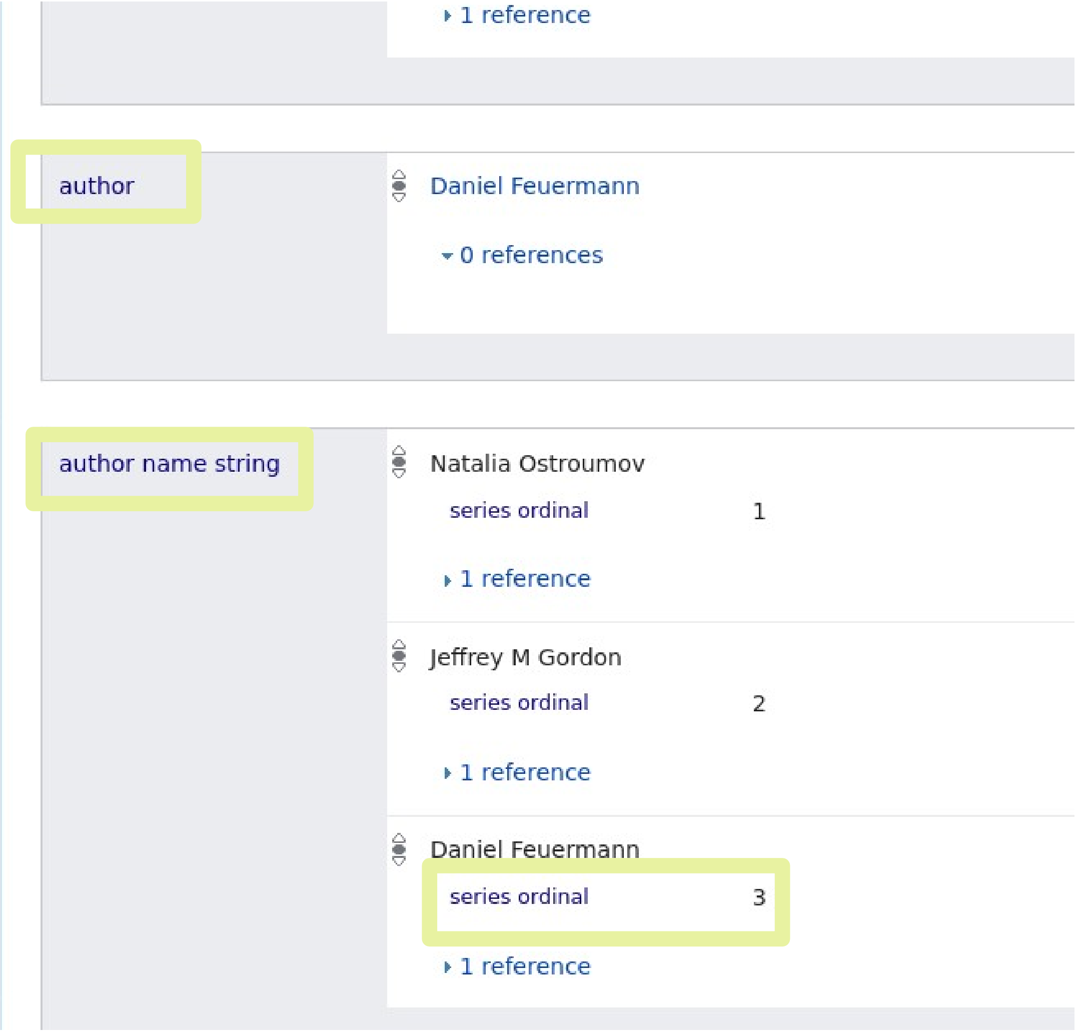
Results: ORCID for Wikidata

ORCID data set 2019

- **Wikidata API approach**
  - 948k author/publication pairs checked
  - 792k authors are not registered in Wikidata
  - 47k author are correct registered to the publication item in Wikidata
  - **>12k authors had been registered as originators to their publication item**
- **Data dump approach**
  - >7.6M author/publication pairs to check
  - 15%: **>33k authors had been added to their publication**

How we continue

- Goal: continous improvement of the existing data
- More information on affiliation, funding... could be imported to the author items
- Missing author items for publications could be created
- Transfer of “series ordinal“ (P1545) from “author name string“ (P2093)
  - Removal of P2093 to avoid confusion of external tools as Scholia



1 reference

author

Daniel Feuermann

0 references

author name string

Natalia Ostroumov

series ordinal 1

1 reference

Jeffrey M Gordon

series ordinal 2

1 reference

Daniel Feuermann

series ordinal 3

1 reference

How we continue

- See the code at GitHub: https://github.com/EvaSeidlmayer/orcid-for-wikidata
- Eva Seidlmayer, Jakob Voß, Tetyana Melnychuk, Lukas Galke, Klaus Tochtermann, Carsten Schultz and Konrad U. Förstner: ORCID for Wikidata. Data enrichment for scientometric applications, https://wikidataworkshop.github.io/papers/Wikidata_Workshop_2020_paper_9.pdf.

Post scriptum

FELLOW PROGRAMM FREIES WISSEN

- Questions?
- Thanks to:
  - Wikimedia and the Fellowship “Free Knowledge”, especially: Dr. Jakob Voß (VGZ)
  - My working group “Data Science and Services” at ZB MED – Information Centre for Life Sciences, especially: Prof. Konrad U. Förstner
  - My Q-Aktiv team: Lukas Galke (ZBW) and Tetyana Melnychuk (CAU Kiel)