In the past, we often discussed how helpful ORCIDs are for our work. An ORCID (Open Researcher and Contributor ID) is a unique personal identifier that scientists can attach to their work. The ORCID ensures that this work is linked to the correct scientist an not to someone else with the same or similar name. We at dblp use ORCIDs to create clean bibliographies. A bibliography should list the work of a single researcher and of course a unique identifier is very helpful here. In this post I will give a short overview on how we handle ORCID and how prevalent it is in DBLP just now. If you do not have an ORCID, consider getting one (for free) at orcid.org. Please make sure that it is attached to your publications whenever possible.

We started experimenting with ORCID in 2016. A more complex integration began in 2017 when we also started to show ORCIDs in bibliographies and individual publications. At the same time we made ORCIDs available with our data releases. We obtain most ORCIDs directly from the publishers together with other publication meta data such as title and author names. ORCID was established in 2012 and many publishers started to attach ORCIDs to their publications only recently (or do not do that at all). But authors can claim such works on their own. This information is provided by ORCID via their annual data dump which we also map to our data set. This means that ORCID has become a common type of data in our collection. Below you see the fraction of signatures in dblp for which an ORCID is known. A signature is a pair of author name and paper. So a paper with five authors has five signatures.
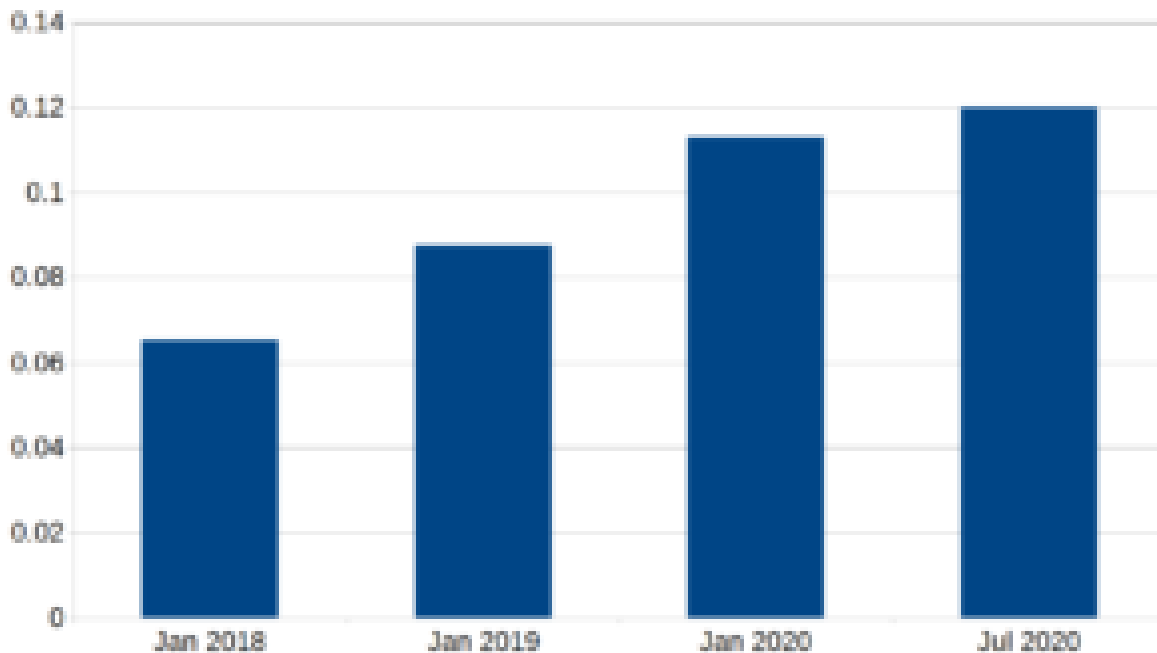
Fraction of signatures with ORCID

An ORCID is now available for 12% of all our signatures and that number is going up. At the moment, we add ORCIDs to dblp in batches. This means that a publication can appear in dblp without any ORCIDs. A few days later they are added. We are working to streamline this process for a faster integration.

Of course signatures from recent publications have a better chance to have an ORCID. However, via the claim mechanism authors can attach ORCIDs to older publications as well. Below you can see the fraction of signatures with an ORCID by the year in which the paper was published:

Fraction of signatures with ORCID by publication year
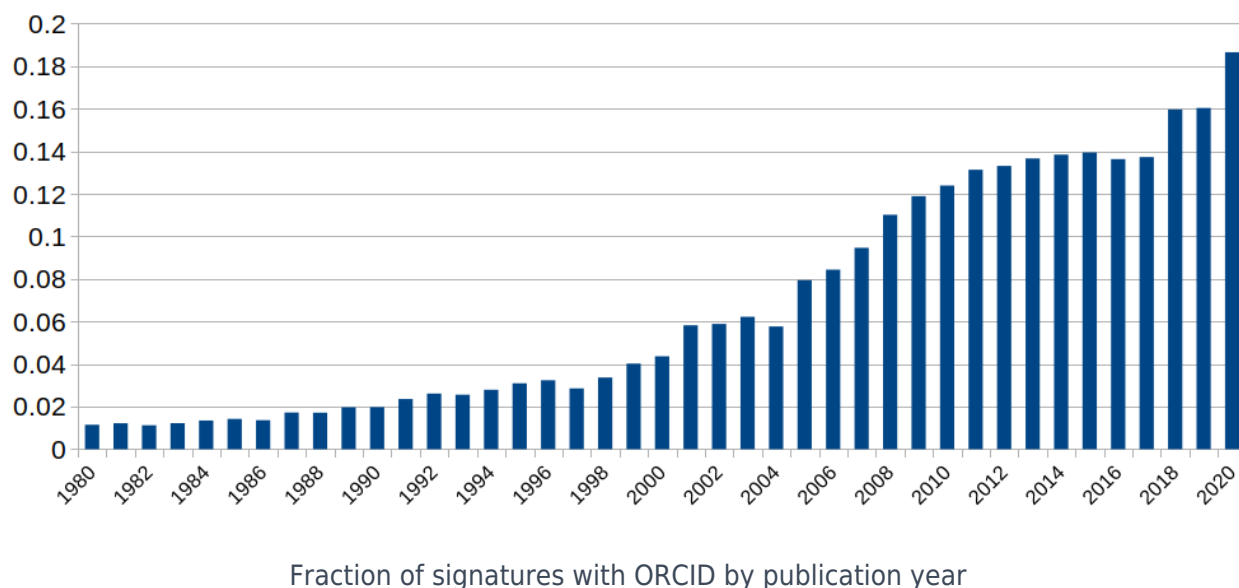
For 2020 we observe a coverage of above 18%. However, even before the year 2000, coverage is above 2% ! The oldest publication in dblp with an ORCID is from 1961. This means that ORCIDs can help with cleaning up bibliographies from a time where other meta data were often problematic (e.g., many abbreviated first names).

The primary reason we use ORCIDs is to create bibliographies. This happens in two ways:

- We find defective assignments in dblp. E.g. a bibliography is associated with two ORCIDs. This probably means that this bibliography actually represents two persons. ORCID plays a major role in the increased number of corrections we discussed here.
- When adding a new publication (we do this manually with support from algorithms) a known ORICID can help to identify bibliographies. E.g., we immediately see that a paper by 'John Doe' should by assigned to the bibliography 'John Doe 0042'. This reduces assignment errors and speeds up the integration process.

As part of this process, we also confirm ORCIDs for bibliographies. You can identify these pages by the green ORCID displayed next to the name (instead of a gray one). We assume that bibliographies with a confirmed ORCID are clean. At the moment there are

69,620 bibliographies with confirmed ORCID. Below you can see how the number of bibliographies with confirmed ORCID developed:
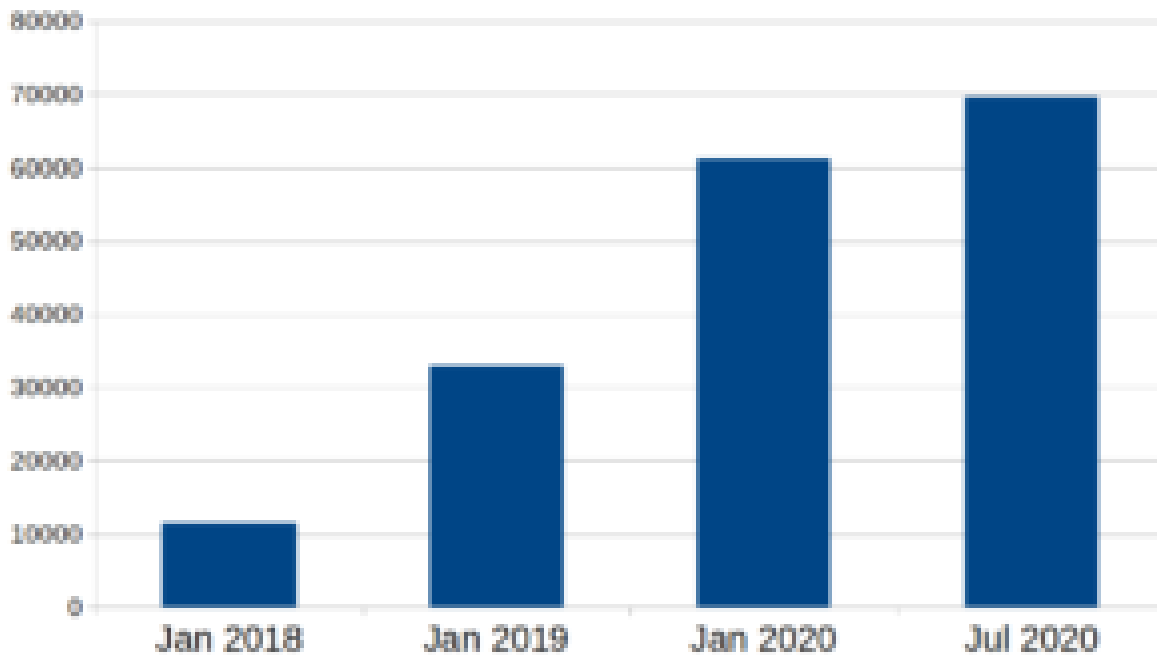
Number of bibliographies with confirmed ORCID

This means that about 2.4% of all our bibliographies have a confirmed ORCID.

ORCIDs are not only used to clean up bibliographies. We also use it to link dblp to other projects. E.g., if there a dblp bibliography with a confirmed ORCID and a Wikidata entity with the same ORCID we link the two entities. At the moment there are 41,121 links to Wikidata in dblp. Many of those are created by matching ORCIDs.

Some random facts

- The International Conference of Applied Computing to Support Industry: Innovation and Technology 2019 (https://dblp.org/db/conf/acrit/acrit2019.html) has an ORCID available for 114 of its 123 signatures. This makes it the best covered proceedings in dblp.
- The IEEE Control Systems Letters is the journal with the highest ORCID coverage (1226 of 1633 signatures or 75%).
- At the moment of this writing Mohamed-Slim Alouini from KAUS has 1203

publications with an ORCID attached to it. No other dblp bibliography has more publications with ORICDs.