Author Items in Wikidata

By Simon Cobb (User:Sic19)

Presented at the WikiCite Virtual Conference on 26 October 2020.

Introduction

As of 11th October 2020, there are some 38.7 million scholarly articles in the Wikidata database.[1] In total, there are, however, only 19.45 million statements connecting works to their authors, with a further 133 million author name strings acting as a placeholder until such linkages can be formed.[2] Less than fifteen percent (14.62%) of work-author relationships have, therefore, been curated. Further, author items are primarily created during bot imports of scholarly articles and, as a consequence, can be left sparse, containing only generic information, such as instance of human and occupation is researcher claims, for example.

This paper provides an overview of the author items associated with the scholarly articles, which have been created in Wikidata as part of the WikiCite initiative. It draws on a combination of data analysis and experience gained by the author whilst importing to Wikidata employment and education data from ORCID records. The paper begins by defining the study sample and then moves on to analyse author items data quality. Next, the conceptual challenges that hinder data import to address deficits are considered before concluding with suggested next steps to improve the situation.

Methodology

This paper is based on analysis of a custom Wikidata RDF dump created using the WDumper tool on 3 October 2020.[3] The data is an extract from the 28 September 2020 Wikidata dump and contains 1,625,666 entities with an ORCID ID (property P496) claim. Due to hardware limitations and time constraints, it was not considered a viable option to work with a larger dataset, which would have provided clarity concerning the total number of authors currently linked to scholarly articles. There is not, however, any indication that a proportionally significant cohort of author items has been excluded from this analysis as a result of the selection criteria.

For purposes of comparison of data held in Wikidata items and ORCID profiles, a ten percent sample (162,224 Wikidata items) was selected using the ORCID checksums starting with 0 (i.e. 0xx).[4] Data about these entities were retrieved from the ORCID public API (https://pub.orcid.org/).

The findings presented below refer to the ten percent sample unless otherwise stated.

[1] Count of instances of scholarly article or subclass of scholarly article: https://w.wiki/gPq.
[2] Count of author (P50), author name string (P2093) and instance of human (P31 Q5) statements: https://w.wiki/gQU.
[3] https://wdumps.toolforge.org/dump/780
[4] https://w.wiki/iSb

Author items data quality

The items studied have an average of 4.67 statements. It is inherent that an ORCID ID (P496) is amongst these statements and, as Table 1 below shows, instance of (P31) is also ubiquitous. Since occupation (P106) claims are found on 80 percent of items, this average can be considered indicative of the omission of basic data that should be stored about humans who are publishing their work in academic journals.

Table 1: Number of Wikidata items with a property claim

	With claim	With claim %	No claim	No claim %
P21 – sex or gender	24841	15.31%	137383	84.69%
P27 – country of citizenship	2749	1.69%	159475	98.31%
P31 – instance of	162224	100.00%	0	
P69 – educated at	15686	9.67%	146538	90.33%
P101 – field of work	621	0.38%	161603	99.62%
P106 – occupation	129696	79.95%	32528	20.05%
P108 – employer	42248	26.04%	119976	73.96%
P569 – date of birth	3982	2.45%	158242	97.55%
P570 – date of death	40	0.02%	162184	99.98%
P734 – family name	15783	9.73%	146441	90.27%
P735 – given name	43135	26.59%	119089	73.41%
P1412 – languages spoken, written…	1006	0.62%	161218	99.38%
P1416 – affiliation	444	0.27%	161780	99.73%

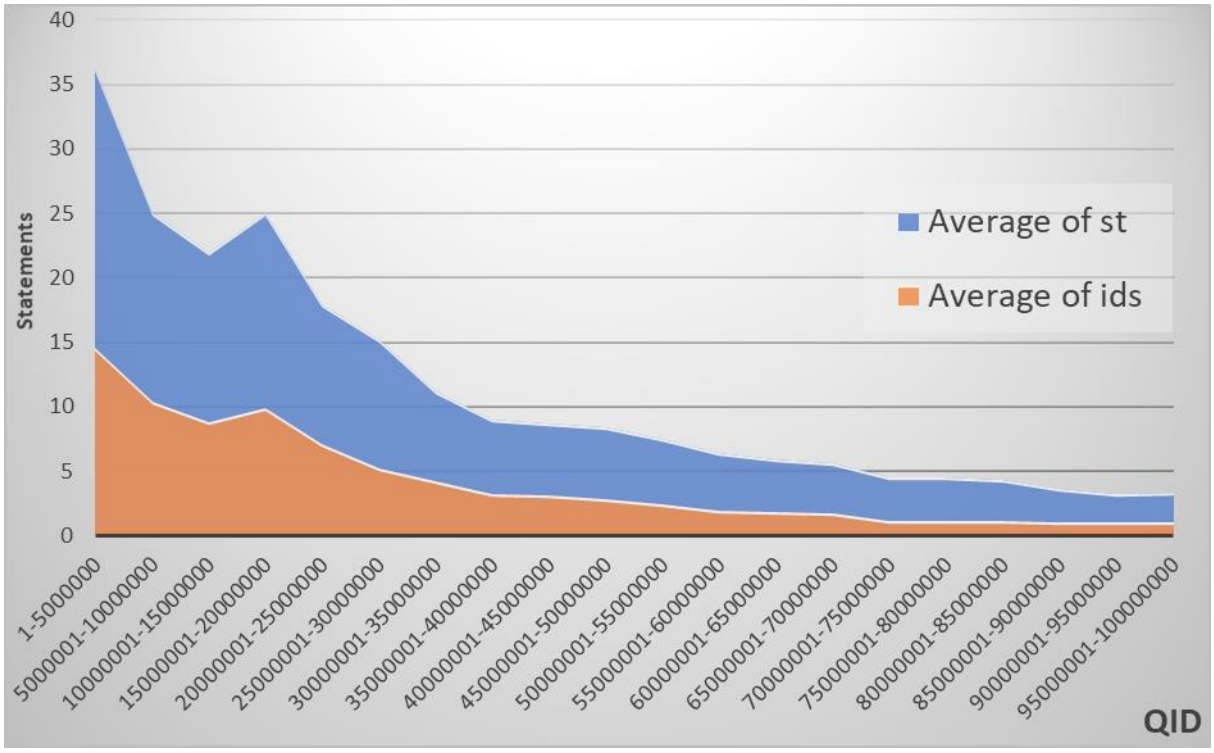

Figure 1: Average number of external identifier properties (orange) and total statements (blue) per QID range

Surname (P734) and forename (P735) data, for example, should be routinely available about the majority of entities with an ORCID. In the same vein, employer (P108) or affiliation (P1416) data can be extracted from research output items, such as scholarly articles, which are linked to author items by author claims. Scholarly articles also provide a data source from which other claims, including field of work (P101) and languages written (P1412), can be extracted or inferenced but, nonetheless, these data are found in less than one percent of items.

As Figure 1 shows, the average number of statements for an author item decrease over time and therefore the most recent creations tend to have the least statements. Whilst it should not be surprising that the latest author items are sparse in comparison to older items, which have had longer to attract the curatorial efforts of community members, there are concerning aspects, nevertheless. For example, items in the Q70000001 to Q75000000 average 4.49 statements despite being approximately one year old (having been created in October or November 2019). The implication is author items created in the past twelve months are, generally, sparse and is especially pertinent because 110,466 (68.09%) of the items sampled were created during this period.

The sparsity of author items added to Wikidata since October 2019 is indicative of bot editing activity which is creating items with very minimal data, i.e. ORCID ID, instance of human, and an English language label. Occupation is research claims are usually added promptly by the batch editing of community members. Whilst there appears to have been an acceleration in author item creation, a corresponding strategy to achieve a level of data quality that would give these items a purpose beyond mere connecting nodes between publications seems to be lacking. It should not be assumed that the creation of sparse author items is sufficient to trigger further curation by Wikidata editors. Indeed, any such assumption is questionable when one considers the difficulty in identifying a subset of items aligned to particular editing foci, such as a research discipline or institutional affiliation, without the relevant connections in Wikidata.

Moreover, there is evidence that author items are being created by bots without even basic validation of data imported. There are, for example, in the data sample items with an ORCID ID that is not nineteen characters in length, as required by the specified format as regular expression claim for this property.[5] It is clear that such errors can persist for many months without being rectified and can be replicated in bulk editing of the description without detection.[6] Likewise, some of the items created would be obvious constraint violations if they were not given generic claims instance of human and occupation is researcher claims. It should not be controversial to suggest that neither the French Society of Pediatric Hematology and Immunology (see Figure 2) nor French Vasculitis Study Group are not human researchers.[7] Both appear to be instances of an ORCID being extracted from a published work and associated with the wrong entity in a list of authors.

There are also issues relating to the latency of data imported and its maintenance. Some 230 items have been identified which have an ORCID that is either deactivated or deprecated (198 created in the previous 12 months). For the latter, ORCID points to another record for the same

[5] 14 characters: Q90804537, Q91303234, Q91522540, Q92758306, Q92758349; 18 characters: Q90247105, Q95856278, Q96128121.
[6] Q90247105 was created on 12 April 2020 and the ORCID ID is erroneous at a glance (0000-0000-000-000X).
[7] See https://www.wikidata.org/w/index.php?title=Q79331041&oldid=1267598446 and https://www.wikidata.org/w/index.php?title=Q82693014&oldid=1266803717.

individual. However, preferred (six items) or deprecated (ten items) ranks are seldom deployed on ORCID claims. This could indicate that Wikidata does not have the latest information to identify these authors and, in some instances, this has resulted in duplicate items being created.[8] In addition to importing data, we must have a plan to maintain it, with periodic checks and updates.

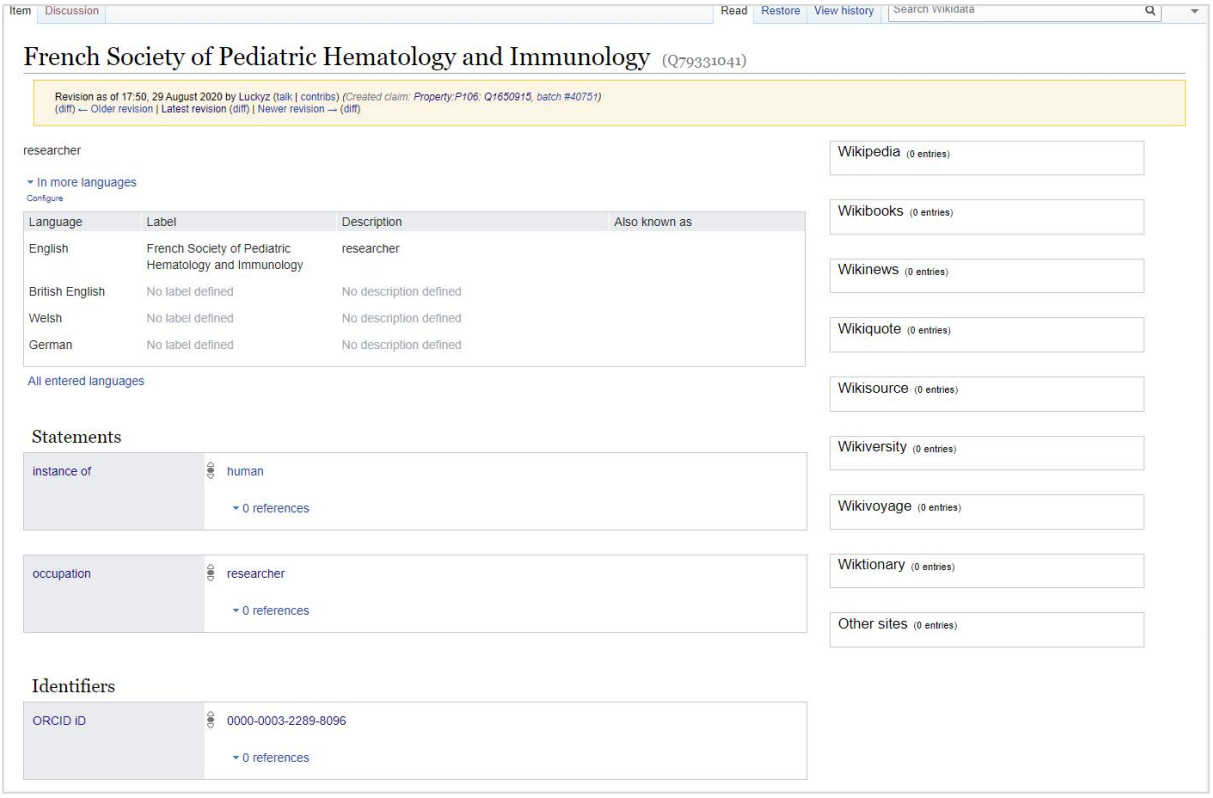

Figure 2: Antoine Benard's Wikidata item (Q79331041) was incorrectly labelled as French Society of Pediatric Hematology and Immunology for ten months.

Taken together, the problems with the completeness, accuracy and latency of author data can be reduced to a simple question; is the data quality of author items in Wikidata considered acceptable? It must be noted that some of the issues identified will be time-consuming for individual editors to unpick and resolve.

In the next section, the challenges encountered when importing data from ORCID records will be discussed.

Data import from ORCID

ORCID provides authoritative, albeit self-curated, data about researchers and, since it is open data, it is ideal for importing to Wikidata. However, the statistics available from ORCID[9] tell us that of the 9,868,477 live records:

[8] For example, see Q262354; https://orcid.org/0000-0002-2662-3092 is deprecated and superseded by https://orcid.org/0000-0002-2658-330X, consequentially, a duplicate item (Q96238990) was created for the author: https://www.wikidata.org/w/index.php?title=Q96238990&oldid=1291006358

[9] Retrieved from https://orcid.org/statistics on 25 October 2020.

- 2,720,936 (27.57%) have employment data.
- 2,719,981 (27.56%) include education data.
- 2,596,322 (26.30%) contain works.
- 4,170,089 (42.25%) have an external identifier for the person, affiliated organisation, funding, work or peer review work.

Whilst these statistics do not reveal how many records contain data or are publicly accessible, it is apparent that we cannot expect to find reusable data in every ORCID. In addition, there are conceptual challenges and other issues that complicate data reuse.

Reconciliation

When employment or education data are available from ORCID, reconciliation with Wikidata is prerequisite of any data import to augment an author item. Three external identifiers are included in ORCID records and facilitate easy reconciliation when they are also attached to the corresponding item in Wikidata. However, as can be seen in Figure 3, some 11,119 (40%) of the

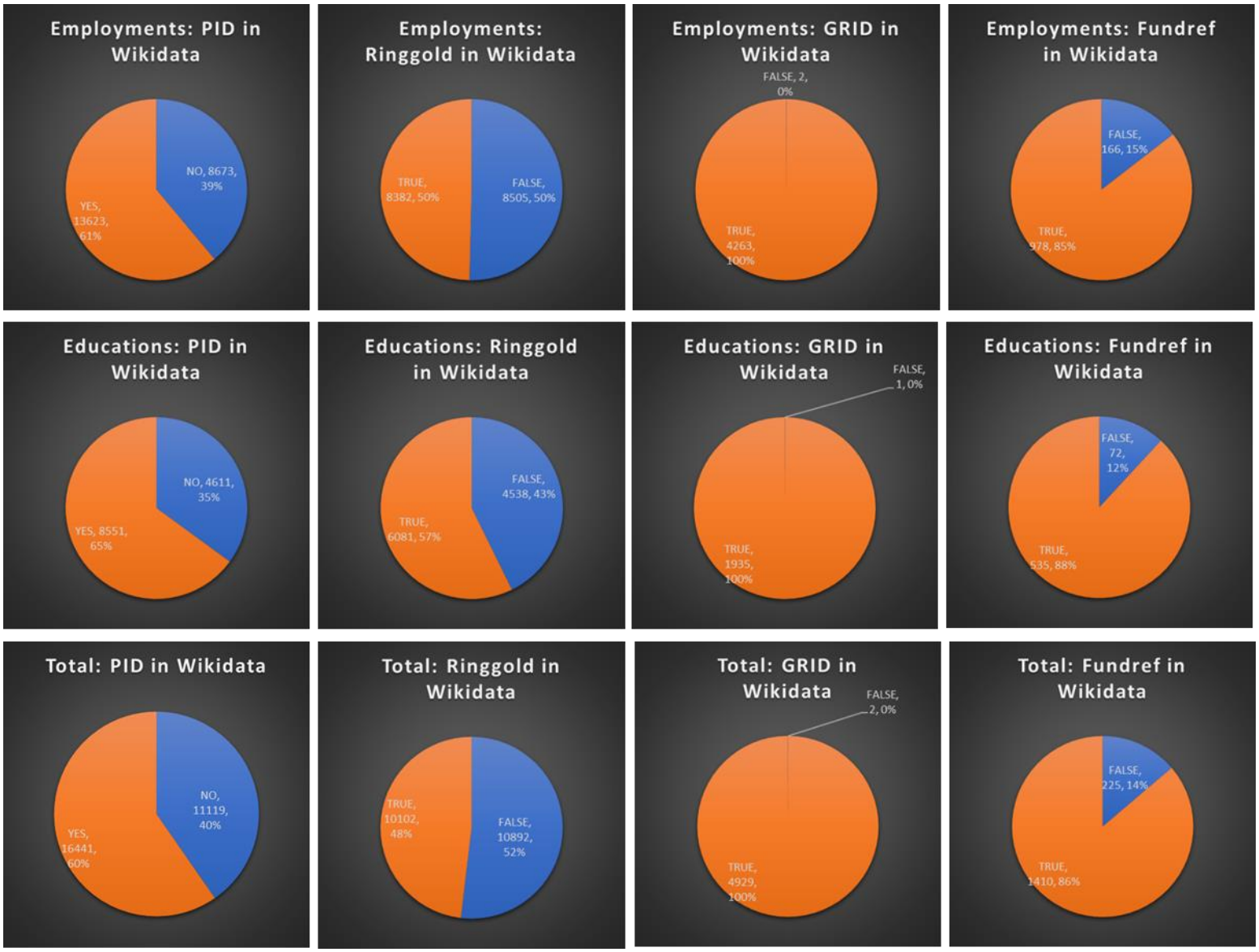

Figure 3: Comparison of external identifiers for organisations in ORCID and Wikidata.

external identifiers in ORCID are not in Wikidata. The lack of these persistent identifiers in Wikidata does not necessarily imply that the organisation does not yet have a Wikidata item; merely that the identifier is not associated with an item.

The problem of missing identifiers in Wikidata is primarily related to Ringgold IDs. Currently, there are 63,639 Ringgold identifiers in Wikidata;[10] almost 500,000 were extracted from ORCID in 2019 by Delpeuch.[11] A substantial amount of importing and curating data about institutions in Wikidata could be required to arrive at an alignment with ORCID than enables mass import of employer and education statements.

Without an external identifier to unambiguously match the organisation in ORCID affiliation data to a Wikidata item, the reconciliation process becomes heavily reliant upon text matching, which is fraught problems due to multilingual text in both the source and target dataset and the myriad of name variants found for an organisation. The problem is compounded by the lack of additional data to perform a validation of matches with sufficient rigour to avoid errors. For example, confirming that a reconciled organisation is in the appropriate country and city is insufficient to ensure that the correct organisation has been matched.

Furthermore, it should be noted that over-reliance on the persistent identifiers currently available in Wikidata could result in bias in Wikidata due to the uneven distribution of organisations with a identifier from the GRID (P2427), FundRef (P3153) or Ringgold (P3500) databases, which is

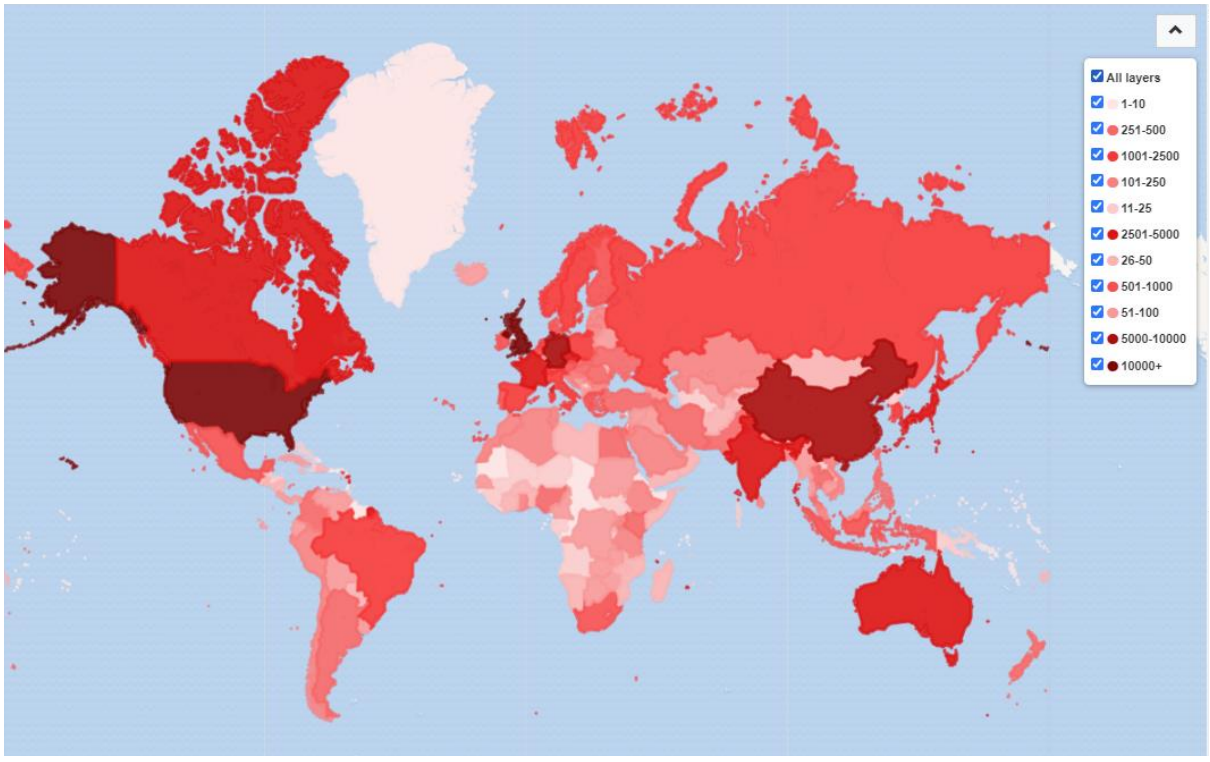

Figure 4: Institutions per country with Ringgold, GRID or FundRef Identifiers (https://w.wiki/iRn)

[10] On 25 October 2020 per https://w.wiki/igf
[11] See Delpeuch, A. (2019) *Aligned ISNI and Ringgold identifiers for institutions*, Figshare. Available at: https://figshare.com/articles/Aligned_ISNI_and_Ringgold_identifiers_for_institutions/8246747 [Accessed 25 October 2020].

displayed in Figure 4. While this may be an interesting area for detailed analysis, it is not within the scope of this paper.

At present, an important aspect of problems related to reconciliation is the difficulty in traversal from a sub-unit of an organisation, such as a faculty or department at a university, to the top-level of the organisational structure. With a viable route to traverse to the university when reconciliation against a faculty, for example, is unsuccessful it would be possible to increase the quantity of data imported.

Overcoming the reconciliation challenges is likely to be key to scaling the import of affiliation data. Other known issues which hinder data imports are detailed below.

Problems in Wikidata

Incomplete dataset

As discussed above, Wikidata does not have an item for every institution and each of its sub-units. Likewise, many occupations, positions or role, academic ranks, fields of research, degrees and other qualifications are not yet in Wikidata. All can be included in employment or education data imports, as appropriate, when reconciliation is successful.

Deprecated identifiers

The handling of withdrawn, redirected, deactivated and deprecated external identifiers is inconsistent. In these scenarios, it is desirable that identifiers remain available for reconciliation with Wikidata items.

Errors

There are persistent identifiers associated with the incorrect entity in Wikidata and consequentially, errors occur in data imports that utilise these identifiers for reconciliation.

A facet of this problem that causes further complications is additional incorrect identifiers being imported due to the relationship with an initial erroneous value. This is a known issue with Ringgold identifiers added using existing ISNI values.

Problems in ORCID

Errors

Occasional errors in ORCID affiliation data result in an incorrect external identifier being associated with an employment or education entry. In Figure 5, the erroneous Universidade de Santo Amaro identifier in a Universidade de São Paulo employment is highlighted.[12] Several hundred instances of this particular error were identified after data import errors were reported to the current author.

[12] The screenshot is taken from https://orcid.org/0000-0001-6801-3519. A similar error can be found in both the employment and education sections of https://orcid.org/0000-0003-2711-1627.

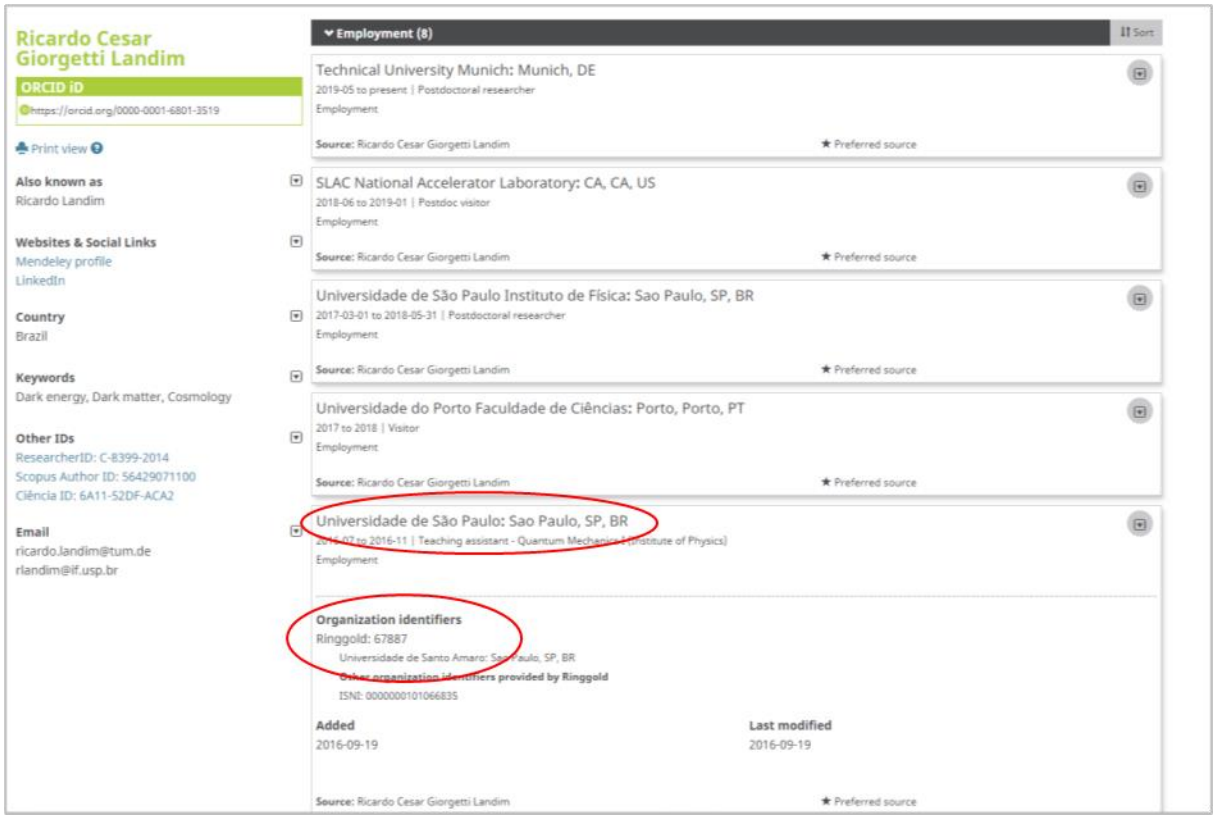

Figure 5: ORCID record for Ricardo Cesar Giorgetti Landim (0000-0001-6801-3519). NB. Ringgold identifier for Universidade de Santo Amaro is erroneously stored in an employment record for Universidade de São Paulo.

Inconsistent data

The inclusion in an ORCID of data entered by both the researcher and institutional account of their employer can result in superfluous or sometime contradictory data being created.[13] This would be replicated if all employment data were imported to Wikidata without prior review.

Granularity

ORCID can be a rich source of employment data, with an entry containing details of the organisation, department, role and dates for each appointment. It can, however, also become excessively granular, such as when multiple entries for a single spell of employment within an organisation are added and no obvious career progression is captured (see Figure 7 for example).

Conflation

Records in ORCID that conflate multiple stages of a researcher's education can cause complications during reconciliation with Wikidata. For example, Figure 6 illustrates the combination of two degrees in a single education entry.

[13] See https://orcid.org/0000-0001-6102-9075 for an example of a superfluous employment entry added by the employer; all of the data was already included before the input from the institutional account. Instances of contradictory start and end dates, input respectively by the researcher and institution have been encountered during work with ORCID data.

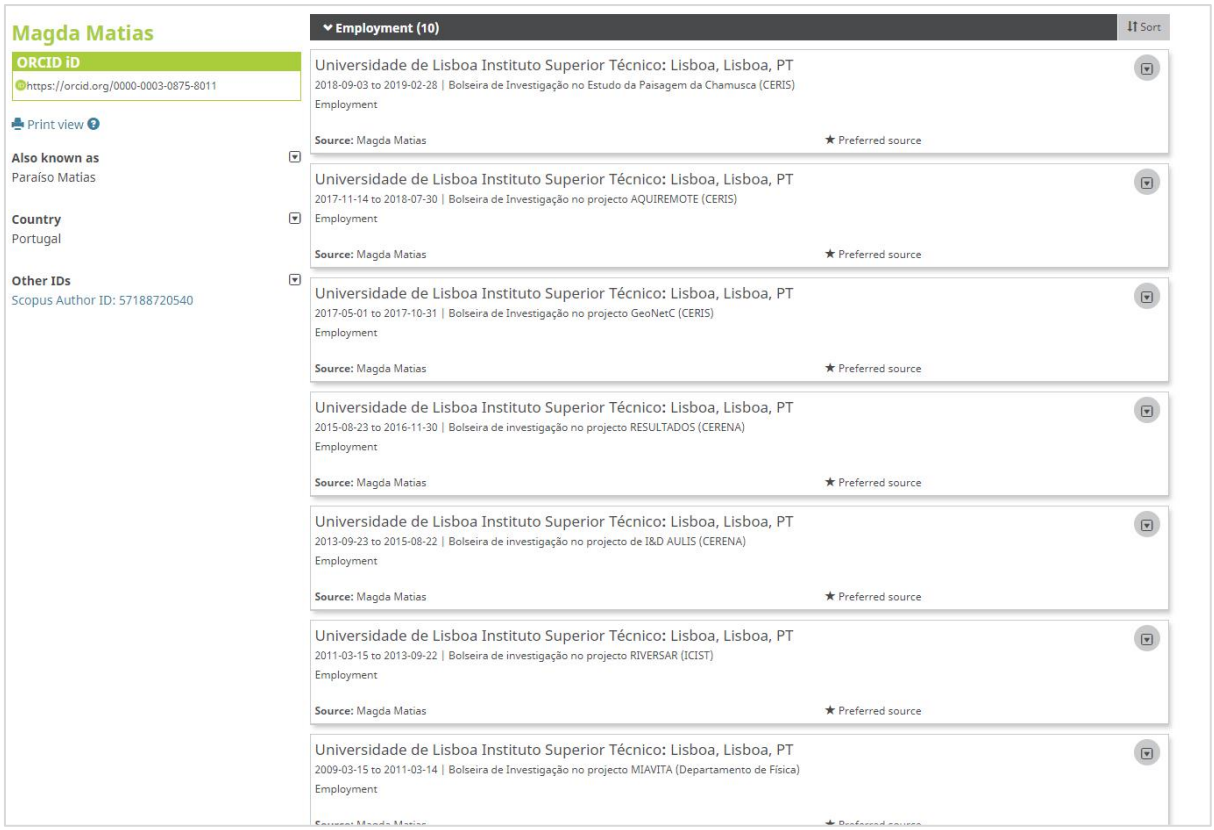

Figure 7: ORCID record for Magda Matias (0000-0003-0875-8011) has a continuous spell of employment from 2010 to present split across eight entries.

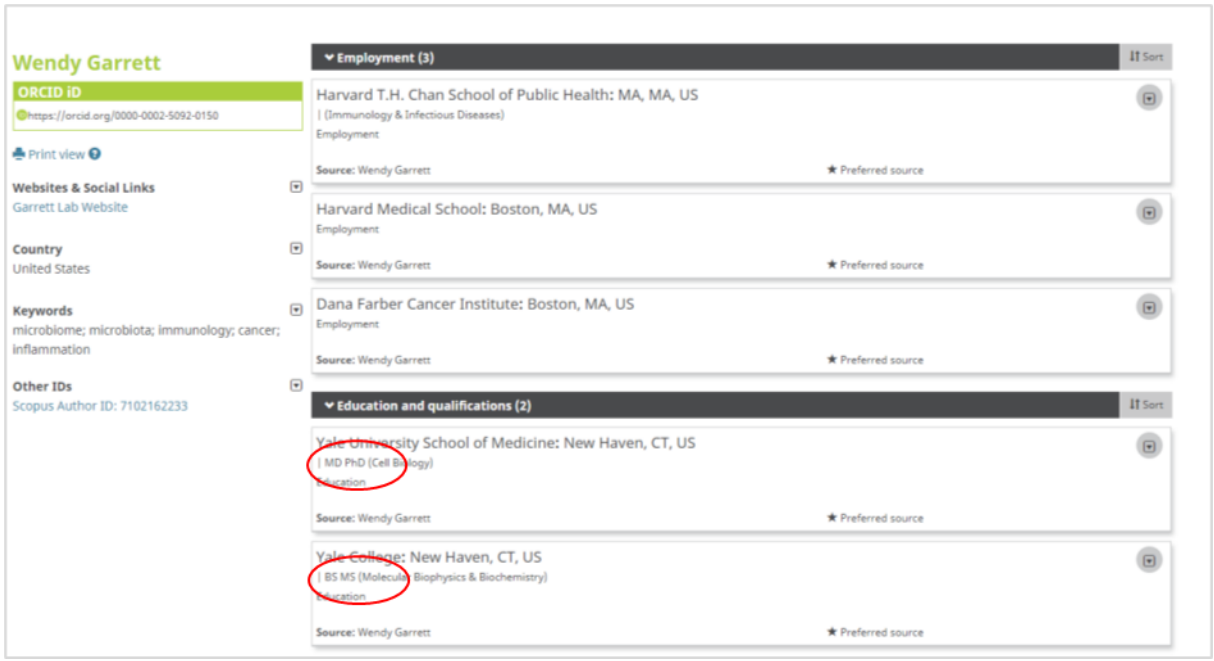

Figure 6: ORCID record for Wendy Garrett (0000-0002-5092-0150). NB highlighted education entries with two degrees are combined.

Similarly, ORCID records which combine, for example, employment, membership and professional activities in employment entries can result in the introduction of errors or nonsensical data to Wikidata.[14]

Validation

Any validation of data imported from ORCID is complicated unless the put code is stored in Wikidata. With the put code, an individual employment or education summary can be retrieved instead of having to fetch all summaries for a researcher using just the ORCID ID.

Suggested next steps

- Seek community consensus on minimum acceptable standard for author items created by bot imports.
- Define author data requirements for a variety of use cases.
- Review and update the FundRef identifiers in Wikidata to ensure complete coverage.
- Identify and remove incorrect Ringgold IDs from Wikidata.
- Continue importing Ringgold identifiers to existing Wikidata items.
- Review and validate data in existing author items.
- Evaluate options for storing put codes from ORCID in Wikidata reference statements to support the validation and updating of data.
- Organise an online workshop to facilitate discussion and collaboration between interested members of the Wikidata editor community and other stakeholders within and outside the Wikimedia Foundation projects.
- Establish a WikiProject or special interest group (SIG) to focus on the improvement and maintenance of author items.

Acknowledgement

This paper is an output of a WikiCite e-scholarship received by the author (for details, see https://meta.wikimedia.org/wiki/WikiCite/e-scholarship/Sic19).

[14] For example, see the employment section of https://orcid.org/0000-0003-2373-2004, which includes appointments, editorships, memberships and other professional activities amongst the 77 entries.