Supplementary Material 10: Further analysis of manual cross-validation results

(see Brookes-Kenworthy et al, 2019, for a summary and raw forms of data collected via the manual cross-validation process).

Figure S10.1: Percentage of exclusive DOIs from each source that are found in the metadata (matching DOI only) of the other two sources.

Here we proceed to validate the sources against each other, at the institutional level. The three charts in Figure S10.1 show the percentages of sampled DOIs from one source that are found in the metadata of the other two sources (i.e., a "DOI match"). This presents some indication of the amount of DOIs that are actually covered by another source, but simply did not link to the target affiliation in our data collection and validation processes (either the object did not show up in the affiliation and year search or the object's DOI is not passed through collection process via APIs)[1].

[1] One should also note here that having the target affiliation appearing in metadata is not necessarily the same as the object actually being linked to the target affiliation systematically (e.g., API searchable).

The top chart indicates that a relatively high proportion of DOIs from WoS are indexed in both Scopus and MSA. This implies that Scopus and MSA actually have good coverages of DOIs that initially seemed exclusive to WoS, but they were not assigned to the target affiliations by Scopus and MSA. In contrast, much less of DOIs from MSA can be found in WoS and Scopus (bottom chart). As for DOIs from Scopus (middle chart), relatively high proportion of them can be found in MSA, but much less so in WoS. An extreme exception is DUT, where neither WoS nor MSA have very high coverage of DOIs exclusively from Scopus (it may be worth recalling we saw earlier that Scopus had exclusive coverage of many DOIs registered with Chinese DOI agencies). Overall, it appears that MSA has a broader coverage of all DOIs, but the completeness of affiliation metadata is lower.

While a DOI may be missing from a database, the object to which the DOI is assigned to may actually still be in the database, i.e., the DOI was simply not recorded in the metadata of the research object. Hence, we also performed “title match” (instead of matching DOI strings) across sources. For each sampled DOI, we check whether the corresponding document title can be found in the other two sources. These are summarised in the three charts of Figure S10.2.

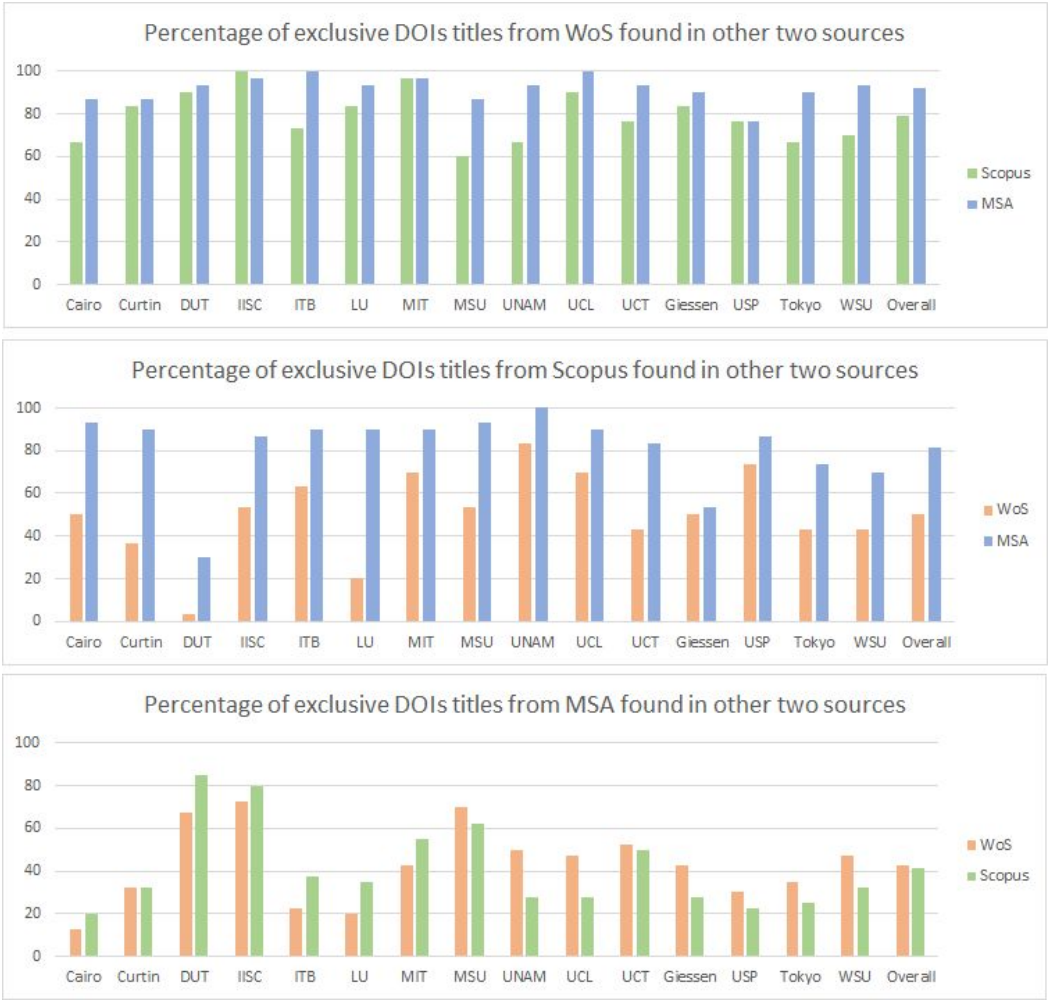

Figure S10.2: Percentage of exclusive DOIs from each source that have their corresponding title found in the other two sources.

All percentages in Figure S10.2 are equal or greater than the corresponding percentages in Figure S10.1. This is because for all cases for which the DOI was found, the corresponding title was also found (i.e., various objects have correct titles but no record of their DOIs, not the other way around). The results here also further highlight the extent of MSA's high coverage of objects related to DOIs initially appeared to be exclusively indexed by the other two sources. Otherwise, the general pattern is similar to what we observed earlier, with WoS titles mostly covered by both MSA and Scopus, MSA having more coverage (than WoS) of Scopus titles, and both Scopus and WoS have relatively lower coverage of titles from MSA.

Having the correct affiliation recorded in metadata is not necessarily the same as having the correct affiliation linkage (e.g., an object may have the correct metadata, but does not show up in the affiliation search). As a way to gain some insight into the degree of this issue, we match the affiliation across sources. When two sources both match a DOI to its target affiliation, we refer this to as an "affiliation match". Figure S10.3 demonstrates the findings again via three different charts. Each bar represents the percentage DOIs from one source having title match and affiliation match with another source. For example, the green bars in the top chart of Figure S10.3 denote percentages of those exclusive WoS titles that were found in Scopus and also have affiliation metadata in Scopus that match the target affiliations.

Figure S10.3: Title and plausible affiliation found in another source.

It certainly appears that more WoS DOIs actually have title and affiliation matches in the other two sources. One can also note the decrease of percentages when compared to Figure S10.2. This is a clear indication that many of these titles were simply not assigned to their target affiliations by the contrasting sources. However, the numbers here also include title matches that does not necessarily have DOI matches. This means we cannot tell how many of these should have been collected from the contrasting sources via our data collection process.

Figure S10.4: DOI and plausible affiliation found in another source.

Hence, we now filter down to objects that have both DOI matches and affiliation matches. These are presented in Figure S10.4 and are indications of numbers of DOIs that our data collection process should have captured (but did not) from each contrasting source, given they have DOIs and are plausibly affiliated to target affiliations. The reason for these to be missing from our collection process is likely[2] to be that the affiliation linkages are broken. This could include various reasons but the most prominent one seems to be that the metadata (as per source website) is not synchronized with the API returns we gathered.

[2] The other reason would be metadata changes between time of data collection and time of manual cross-validation. But given that we are primary using 2016 data, the scale of this is expected to be relatively small. Manual spot checks did not find any cases where the metadata appears to have changed.

More WoS DOIs have both DOI and affiliation match by other sources, in contrast to those of Scopus and MSA.