Identifying corresponding-author papers from Scopus

Andrew Gray - andrew.gray@ucl.ac.uk
Bibliometric Support Officer, University College London
February 2020

Introduction and overview

As costs for publishing papers rise, and proposals such as Plan S introduce the possibility of large-scale switching to paying to publish, including under so-called "transformative" or "transitional" deals, institutions have a substantial interest in estimating the potential spending that could result from these approaches. As the majority of publications are collaborative across institutions, however, some way of apportioning cost needs to be found.

UK institutions generally use corresponding authorship or grant-holding institution to determine a paper's eligibility for APC funding from their UKRI and COAF open access block grants. Emerging transformative deals are seeing publishers working towards enabling the UK to publish 100% of its UK research outputs open access on publication, with UK and individual institutions' outputs determined by corresponding authorship.[1] This is a reasonable convention: unlike first/last author status, the assumption that the corresponding author is primarily responsible for the work is reasonably consistent across different fields.[2]

In order to assess the likely extra costs of a wholesale pay-to-publish model, institutions have begun to develop methodologies for identifying the proportion of articles and conference papers with a corresponding author at their institution. This enables them to estimate the total additional contribution required from UKRI, COAF and institutional funds, and to make a judgement about particular transformative deals, considering (for example):

1. Additional cost of a transformative deal (the "publish" element) as a percentage of an institution's remaining open access funds, compared with percentage of articles published with the relevant publisher.
2. Cost per article, based on the additional cost of a transformative deal (the "publish" element), measured against total number of articles expected to be covered by the deal.
3. Cost per article, based on the total cost of a transformative deal (the "read" and "publish" elements) measured against total number of articles expected to be covered by the deal.

Institutional publications management systems such as Pure or Symplectic Elements do not usually identify corresponding authorship. However, Scopus contains a field, "Correspondence address", which lists the address of a corresponding author, when provided by the publisher. While this field is very useful, it has certain limitations. It cannot be searched, and so analysis needs to be done on downloaded datasets. It is given exactly as provided (including typos!), and is not normalised or reconciled to Scopus's own database of research organisations. Where the corresponding author has two or more affiliations, it appears that only the first one is listed. Individuals with a dual-affiliation in the correspondence address are exceptionally rare, and usually only when it has been written as a single address line.

[1] https://www.jisc-collections.ac.uk/Transformative-OA-Reqs/
[2] It is still common to see first author (or, less often, last author) used for this purpose, but many fields either stick to a purely alphabetical system or use a mix of different conventions.

The following paper outlines a process developed at University College London (UCL), in collaboration with four other institutions,[3] to assess corresponding authorship, using Scopus, in a relatively straightforward and consistent fashion. It is presented as a series of suggested steps that could be followed by other institutions. Many thanks to colleagues who assisted with this methodology and analysis, particularly Stephen Pearson at Manchester for the efficient download process outlined in step two.

After analysis, approximately 40% of UCL papers published in 2017-18 had a UCL corresponding author, with another 5% from one of the other COLIM institutions. The majority of the rest are other UK universities (12%) or international collaborators (35%). Non-university UK bodies made up only a small share of the total, around 8%, predominantly NHS institutes.[4] The share of corresponding versus co-authorship was reasonably consistent across different journals and publishers, though there were some differences – for example, Taylor and Francis papers were more likely to have a UCL corresponding author, as were those in the large open-access megajournals.

Data extraction and analysis

1. Work out a detailed Scopus search for your institution - the one we used for UCL was designed to pull out all separately identified sub-units (e.g. UCL Institute of Child Health), filtered to just show articles/reviews. We avoided specifically searching for the associated NHS trusts – anything jointly published will be picked up, but anything *purely* from the NHS side without another UCL collaborator is out of scope, as we would not generally consider that a "UCL paper". Exactly where to draw this line may vary between institutions, especially since Scopus is not consistent in how it handles and identifies NHS trusts.

   The Scopus hierarchy contains a pre-generated search for "Documents, whole institution" which returns those matched to the main ID plus those matched to any subsidiary units. We recommend using this as the basis for constructing your search, but you may want to consider a text search in the affiliation address as well.

   For several reasons, in our initial analysis we excluded conference papers. These outputs are of most relevance in specific disciplines (e.g. Engineering and Computer Science), they are generally less well-indexed in Scopus than journal articles, and some publishers do not yet have a model for including them in transformative deals. Any analysis of value for money in a transformative deal that includes substantial numbers of conference papers should adjust for this.

[3] The five COLIM institutions – Cambridge, Oxford, Imperial, UCL, and Manchester

[4] Papers without a declared corresponding author were omitted from these calculations.

2. Export all records. This can be quite a complex process, as Scopus only provides basic metadata for batches of more than two thousand items. The best approach is to use the "citation information only" download (covering up to 20,000 items), then extract the EID field.[5] Load this into Scopus in batches of slightly under 2000 records, using the advanced search,[6] and export each of these in order, then paste the files into a single document, checking for any overflowing lines.

   For the export, you will need to add "**correspondence address**" from the "bibliographical information" section. We would also recommend adding "**publisher**" (see section 5) and removing "author(s)", to try and limit overflow problems. (If you are not planning to count authors, in step 5, you can also remove the "author(s) ID" line, for the same reason). Make sure that whatever other fields you export you always include EID, as this will mean you can go back and link in extra Scopus data later, and DOI, which is useful for linking in other datasets.[7]

3. Open the CSV, save in Excel, and add a new column for manual coding. Filter the **correspondence address** column to find any entries which match your institution – for example, for UCL we looked for any that had a "ucl.ac.uk" or "ioe.ac.uk" email address, any that had the "UCL" keyword but didn't come from "UCL Hospitals", any others that had a Gower St or Bedford St address, and so on. If in doubt, we erred on the side of inclusion – someone with a UCL postal address but a non-UCL email, or vice versa, counted as UCL. (There were a surprising number of these, suggesting a high level of dual-affiliated individuals.)

4. Having identified all local authors, complete the corresponding author indexing. We coded the remaining correspondence addresses into "UK universities", "NHS", "other UK", "international". Depending on the analysis, it may be practical to subdivide by individual institution, country, etc, here, or it may be desirable just to distinguish between "my institution" and "everyone else".

   In addition, we found that around 11-12% of papers had a blank correspondence address.[8] We recommend that when analysing the data, these papers are coded as [blank] or similar and omitted from the calculations. They appear to represent a mix of papers with no defined corresponding author, and ones where it was not provided to Scopus in the publisher metadata.

---

[5] A small number of papers have too many authors to be reliably opened in Excel, and cause the lines to overflow. These cases can easily be spotted by sorting/filtering the EID column – they will not contain a Scopus EID in that cell – and picking out the easily distinguished 2-s2.0-… entries from other cells in that row.

[6] These search syntax here is "EID(x) OR EID(y) OR …". It can be generated from a column of EIDs in excel by adding "EID(" in the column before, and ") OR " in the column after – note the space after "OR", then trimming the very last "OR" of the batch when pasted. The search ignores tabs and linebreaks.

[7] The DOI is not a perfect unique identifier – multiple items can have the same DOI, and a single item can have multiple DOIs – but it is the best one we have access to for most purposes.

[8] Missing corresponding authors were very unevenly distributed between different publishers and authors.

5. Optionally, add further manual coding columns. For our analysis, we added:

    i. publisher, to allow us to group different imprints together (eg "Wiley" and "Blackwell" could be matched together. For simplicity, we grouped all publishers with <100 papers as "small publisher".
    ii. author numbers – single-authored papers could be considered our responsibility even if they notionally seemed to be led elsewhere. Author numbers papers can be identified by counting the number of semicolons in the "Author(s) ID" field.[9] While this breaks down for any cell large enough to overflow, it is certainly able to give an indicative figure to distinguish between single-authored papers, collaborative papers, and those with a very large number of authors.

6. It is possible to tie in extra data, matching on the DOI. We have had some success with tying in actual spending data from our open access APC records. Similarly, it was possible to link up publications with a dataset of papers submitted by UCL to ResearchFish as part of UKRI grant reporting.

    We experimented with using the Scopus open-access data provided by the export, but comparison with internal records suggested it is a substantial undercount and should be treated with caution.

We hope that these suggestions are useful to other institutions. For more information on how UCL is using the resulting Scopus data, alongside existing analyses of APC spend by publisher and funder and prior subscription costs, contact catherine.sharp@ucl.ac.uk.

[9] An Excel command to count this is **SUMPRODUCT(LEN(A2)-LEN(SUBSTITUTE(A2,";","")))+1**