Contextual Data for REF 2021

This document describes the contextual citation data that Clarivate Analytics will provide to the REF sub-panels for REF 2021. This proposal has been defined in collaboration with the REF team.

This document accompanies and describes the 'contextual data example' spreadsheet comprising a mock-up of how the contextual citation data will be provided to the REF team for the REF 2021 sub-panels. During the development of the assessment phase systems, the REF team will explore whether the data can also be integrated into this system, to allow its presentation to sub-panel members in additional ways.

As discussed with the REF team, the decision regarding the subject categories and document types that are most appropriate for a given output will be made by the REF sub-panel members.

Rationale for our proposal

In REF 2014 the contextual data provided to sub-panel members were looked upon by as being difficult to navigate. The data were provided as a single, large, alphabetically-ordered table, so finding the relevant contextual information took significant amounts of time. We are keen to make the contextual data for REF 2021 as easy to navigate as possible. To this end we propose three changes to the way in which the data are provided:

- **Ordering**: We propose to order the subject categories so that those that are most relevant to each Unit of Assessment (UoA) will appear at the top of the data tables. To this end, the percentage of matched submissions to a given UoA that are associated with each subject category will be used to indicate relevance. The contextual data provided in the mock-up spreadsheet are therefore ordered by the percentage of outputs submitted to REF 2014 in UoA 1 that were associated with each subject category, as an example. For REF 2021 we would provide a separate sorted table of contextual data for each UoA.
- **Grouping**: We have also grouped the subject categories into two less granular levels of classification to allow for easier filtering of some of the larger tables in an effort to make identifying relevant subject categories more straightforward.
- **Lookup tool**: We have also tried to make the navigation of the contextual data more flexible by providing a look-up facility as a separate worksheet. We still propose to provide tables covering all subject categories, years and document types to all panels – more-or-less as provided for REF2014 – but with ordering specific to each UoA. However, we also propose to provide the ability to look-up specific datapoints by subject category, year and document type using a separate worksheet to reduce the burden of having to navigate a large table. Sub-panel members can choose the mode of working which suits them best.

In addition, we are proposing to provide separate sets of contextual data for each of the three document types that are most likely to be submitted to REF 2021 (i.e. articles, reviews and conference contributions). Document type was not taken into account last time but, as has been demonstrated, it does make a substantial difference to the number of citations an output receives. While this does increase the volume of contextual data, it will allow a more appropriate contextualisation of the citation

data provided to sub-panel members. We hope that the changes above will mitigate against any increase in effort while enabling better decision-making.

Description

Citation data reflect the impact a publication has had on the field to which it relates, although they do not allow a value judgement about the nature of that impact nor do they indicate whether the sentiment of the citing author was positive or negative in nature. However, many bibliometric studies have shown that citation rates are also affected by publication year, field and document type. To allow REF 2021 sub-panel members to account for these factors in their decision-making, the contextual data will be provided to sub-panels by publication year, subject category and document type. For each combination of these factors we provide six values:

- The number of citations a paper would need to be ranked amongst the world's top 1% of most highly-cited papers.
- The number of citations a paper would need to be ranked amongst the world's top 5% of most highly-cited papers.
- The number of citations a paper would need to be ranked amongst the world's top 10% of most highly-cited papers.
- The number of citations a paper would need to be ranked amongst the world's top 25% of most highly-cited papers.
- The number of citations a paper would need to be ranked amongst the world's top 50% of most highly-cited papers.
- The world average (arithmetic mean) number of citations papers have received since publication.

The first five of these values are thresholds which can be used to indicate where the citation count of a given research publication lies in the distribution of citations to all publications in the same year, field and document type. The final value indicates the arithmetic mean number of citations received by papers in same publication year, field and document type.

The data are presented in two forms:

- **Data tables** – Data are provided as three worksheets, each comprising a single table which allow sub-panel members to look up contextual data values for each of the document types (i.e. article, review or conference contribution). This ensures that all sub-panel members have access to all of the contextual citation data that we will provide.
- **Lookup worksheet** – A formula driven worksheet which allows sub-panel members to identify specific publication years, subject categories and document types, and which returns the relevant contextual data values. This is to reduce the effort required by sub-panel members to look up the contextual citation data they require.

In each case the lists of subject categories provided are ordered by relevance to the UoA in question. As described above, the relevance is indicated by the percentage of publications submitted to the REF that are assigned to each subject category.

Usage

This Section of the document describes how the contextual data spreadsheet can be used to identify the relevant contextual data points for a given output.

The data submitted by the HEIs will indicate the publication year, and sub-panel members will have to select the relevant subject category and to assess whether the output is an article, review or conference contribution based on their expert judgement.

Data tables

To use the data tables, a sub-panel member would have to select the worksheet corresponding to the relevant document type (article, review or conference contribution). The sub-panel member would then need to look up the relevant subject category in column C and the relevant publication year in row 1. Sub-panel members can also filter the worksheet to show only selected research areas using the filters for columns A, B and C.

Example: An oncology article published in 2015 that has received 44 citations

In the accompanying mock-up contextual data spreadsheet, a sub-panel member would select the Article worksheet, look up the subject category *Oncology* in column C (row 5) and the publication year in row 1 (columns K to P). The data would indicate that this article has received more than 30 citations (the number of citations required to be ranked among the world's top 10% of papers; cell M5), but fewer than 55 citations (the number of citations required to be ranked among the world's top 5% of papers; cell L5). The sub-panel member could therefore conclude that this article was ranked in the world's top 10% of most highly-cited papers. The data would also indicate that the article had received ten-times the world average number of citations for a paper in the same publication year, field and document type (4.40 citations; cell P5).

Lookup worksheet

To use the lookup worksheet, a sub-panel member would have to select the relevant document type in cell C3, the relevant publication year in cell C4, and the relevant subject category(ies) in cells B8 to B17. This would then populate the table (cells B7 to I17) with the relevant contextual data. The sub-panel member could (optionally) indicate the citation count received by an output in cell C5.

Example: A review published in 2017 relating to haematology and virology that has received 67 citations

In the accompanying mock-up contextual data spreadsheet, a sub-panel member would select the document type review in call C3, the publication year 2017 in cell C4, and the subject categories *Haematology* and *Virology* in cells B8 and B9, respectively. The sub-panel member could also enter the citation count of 67 in cell C5. The data would indicate that the review had received more than 39 citations (the number of citations required to be ranked among the world's top 25% of papers in both subject categories; cells F8 and F9). However, the review would have received fewer than 70 citations (the threshold required to rank among the world's top 10% of papers in *Haematology*) but more than 65 citations (the threshold required to rank among the world's top 10% of papers in *Virology*). The sub-panel member could therefore conclude that this review was ranked among the world's top 25% of reviews in *Haematology* but among the world top 10% of reviews in *Virology*.