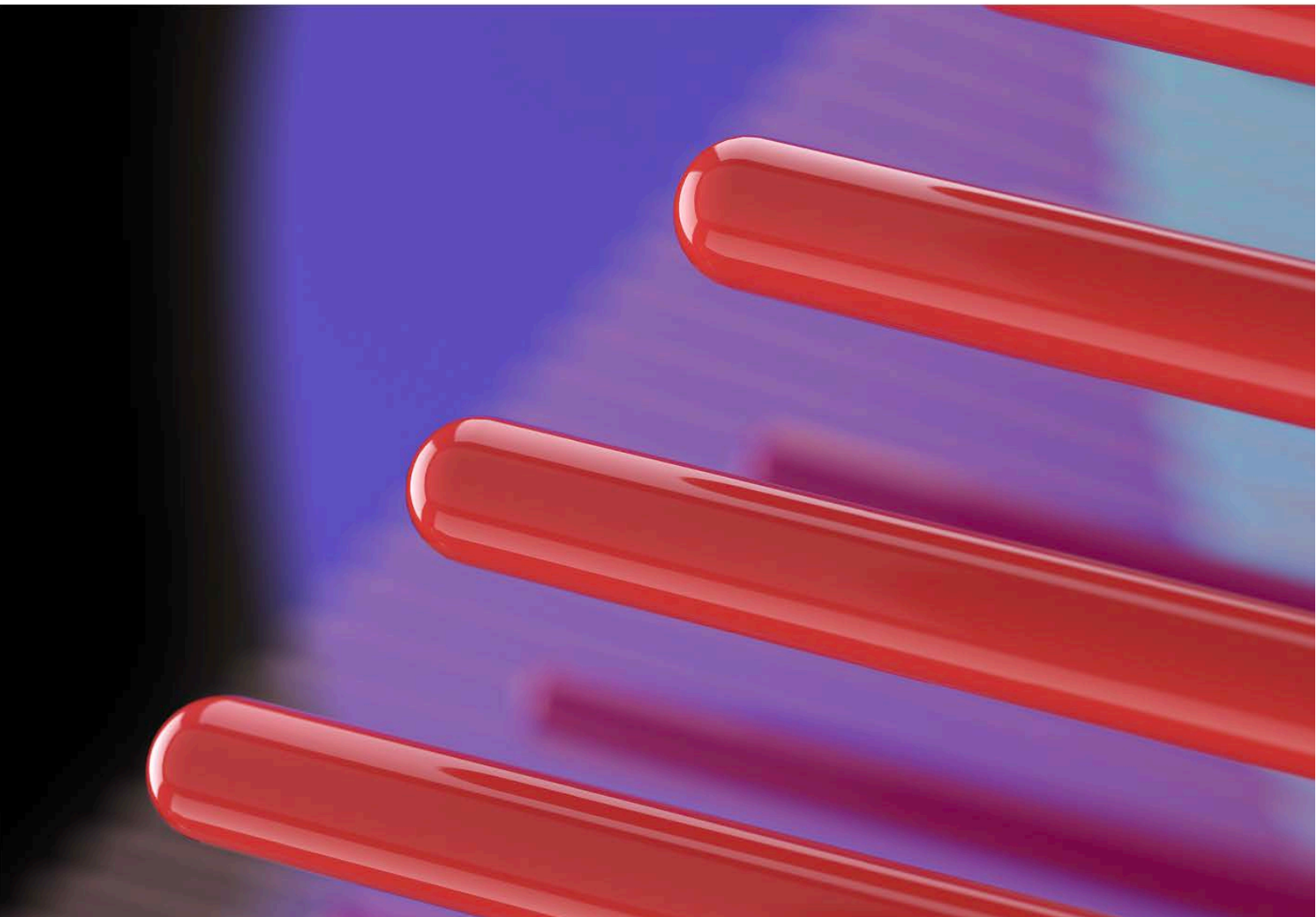
Web of Science Group

A Clarivate Analytics company

Specification

Contextual data for REF 2021

01.07.19



This document specifies the contextual citation data that Web of Science Group will provide to the Research Excellence Framework (REF) team for use in REF 2021. In addition to contextual citation data, Web of Science Group is providing a matching Application Programming Interface (API) to the REF team. The content covered by the API and the contextual data will be the same.

Why do citation counts require contextualisation?

A publication accumulates citation counts when it is referred to by more recent publications. There are many reasons why one publication cites an earlier one, but in general it indicates an intellectual contribution of the earlier publication to subsequent research. Because of this, and on the basis of studies linking citation indicators and peer judgments, it is generally accepted that higher average citation counts correlate with the academic impact of research publications. However, citation counts are also affected by other factors:

- Publication and citation rates are dependent on the field of research. For example, some fields such as cell biology publish more frequently than other life sciences.
- Papers in some fields are typically cited over different time periods than in others. Compare artificial intelligence research to medieval history.
- Citation rates vary by document type, with some document types such as reviews typically receiving more citations than articles and proceedings papers.
- A paper published in 2010 has had longer to accumulate citations than one published in 2017, so on average older papers typically have higher citation counts than more recent ones.

It is therefore important to contextualise a publication's citations before inferring anything about its academic impact through comparison to other publications.

Principles

The following principles underpin the development of the specification presented in this document:

1. The role of Web of Science Group in REF 2021 is to provide citation data and advise on the use of its content; it is not to make decisions regarding what data are to be provided nor to determine how those data will be used. All decisions regarding the content of the contextual data and their usage will be made and signed-off by the four UK HE funding bodies considering the views of all relevant stakeholders.
2. This document specifies the format in which Web of Science Group will provide contextual data to the REF team. However, the final decision about how contextual data are provided to Higher Education Institutions (HEIs) and sub-panel members remains with the funding bodies.
3. The use of metrics in the REF will be responsible; it will take into account best practice (as described in publications such as the Metric Tide and the Leiden Manifesto) and the views of the UK Forum for Responsible Research Metrics.
4. Decisions regarding the nature of each output should be made by the sub-panel members rather than by Web of Science Group, the REF team or other parties. Web

of Science Group will, therefore, not assign individual outputs to journal subject categories or document types in the data provided; these decisions will be left to the expert judgement of the sub-panel members.

5. The process of data provision will be open. The REF submissions system will provide HEIs with the relevant bibliometric metadata from Web of Science so that they can confirm if the matches made are correct. The submissions system will also provide HEIs with the citation count for each document. Contextual data will be made available to HEIs by the REF team as a snapshot at the beginning of the pilot and live submission stages of the REF, and a further snapshot will be provided to panels and made available to HEIs at the close of the submission process. The specification for the contextual data (i.e. this document) will be published on the REF website so that it is fully available to HEIs, sub-panels and other stakeholders.
6. The data should be as transparent as possible. We will therefore not provide normalised (adjusted) indicators or other bibliometrics for individual outputs. Instead we will provide citation counts for each output and contextual information that allow those citation counts to be interpreted.
7. Unless there is a good reason for making changes, the data and the format in which they are provided to the REF team will be broadly consistent with those provided for REF2014.

Data content and coverage

The contextual data we provide to the REF will be derived from the Web of Science Core Collection, a curated collection of over 20,000 high-quality scholarly journals published worldwide (including Open Access journals) and more than 150,000 conference proceedings across the natural sciences, social sciences and humanities. Specifically, the data will come from five citation indexes covering peer-reviewed journal publications and conference proceedings:

- Science Citation Index – Expanded.
- Social Sciences Citation Index.
- Arts and Humanities Citation Index.
- Emerging Sources Citation Index – which extends coverage beyond the core high-impact literature. The journals selected for inclusion in this index have been identified by our editorial team as important to researchers, key opinion leaders, funders, and evaluators regionally and worldwide.
- Conference Proceedings Citation Index – covering the sciences and social sciences.

While the Web of Science platform hosts other publication and citation databases, the Web of Science Core Collection provides consistently collected and curated citation data allowing meaningful comparisons and evaluative judgements to be made.

It is worth noting that coverage of computer science conferences has recently been enhanced in response to gaps identified by Web of Science Group[1].

[1] The additional conferences that are now covered are: International Conference on Autonomous Agents and Multiagent Systems; Functional High-Performance Computing; Foundations of Software Engineering;

We will 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 proceedings papers). Document type was not considered separately in REF 2014 but, as has been demonstrated, it does make a substantial difference to the number of citations an output can be expected to receive. 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 will provide separate contextual data points for each of the 254 Web of Science journal subject categories[2]. These subject-based categories cover the natural sciences, the social sciences, and the arts and humanities. They group journals into connected topics based on their primary academic contributions. Papers in multidisciplinary journals are assigned to specific categories on the basis of the papers they cite. The journal categories provide a more granular classification than the Units of Assessment (UOAs) used in the REF, and thus allow citation counts to be contextualised at the sub-discipline level. While Web of Science Group will provide contextual citation data for all journal categories with sufficient publications, only 11 UOAs will actually use citation data to inform their judgements.[3]

Where very few articles, reviews or proceedings papers have been published in a journal category in a particular year, we will not return contextual information. We will decide on the threshold for providing contextual data in consultation with the four UK funding bodies.

The decision regarding the subject categories and document type that are most appropriate for a given output will be made by the REF sub-panel members. This ensures that selection of these parameters is made by subject matter experts and does not rely on algorithmic assignments made by Web of Science Group. It should be noted that more than one subject category may be relevant to the content of a given output and sub-panel members may want to refer to more than one set of contextual data when making their assessment.

We will provide static contextual datasets at three timepoints during the project:

- Start of the submissions system pilot phase (autumn 2019)
- Launch of the submission system (January 2020)
- For the assessment phase (November 2020).

The publication window for outputs submitted to the REF is from 1 January 2014 to 31 December 2020. However, as the final citation data being provided to the sub-panels are being collated during the calendar year 2020, papers published during 2020 are very unlikely to have had sufficient time to accumulate a level of citations that could be used

---

International Conference on Functional Programming; Languages, Compilers, and Tools for Embedded Systems; Advances in Neural Information Processing Systems; Object-Oriented Programming Systems, Languages, and Applications; Partial Evaluation and Program Manipulation; Programming Language Design and Implementation; Principles of Programming Languages; Principles and Practice of Declarative Programming; ACM SIGIR Conference on Research and Development in Information Retrieval; International Conference on Management of Data; Principles of Database Systems; Ubiquitous Computing; and World Wide Web.

[2] The Journal Subject Categories are listed on our website: https://images.webofknowledge.com/images/help/WOS/hp_subject_category_terms_tasca.html.

[3] These UOAs are: 1. Clinical Medicine; 2. Public Health, Health Services and Primary Care; 3. Allied Health Professions, Dentistry, Nursing and Pharmacy; 4. Psychology, Psychiatry and Neuroscience; 5. Biological Sciences; 6. Agriculture, Food and Veterinary Sciences; 7. Earth Systems and Environmental Sciences; 8. Chemistry; 9. Physics; 11. Computer Science and Informatics; and 16. Economics and Econometrics.

reliably to infer their likely academic impact in the medium- to long-term. We will therefore not provide contextual citation data for outputs published in 2020.

We would also advise caution in interpreting citation data for outputs published in 2019. A paper published in that year could have a higher citation count than another simply because it was published early, in January, rather than later in the year. For the two static snapshots of contextual data provided prior to the pilot submissions phase (autumn 2019) and the main submissions phase (January 2020), we will provide contextual citation data for the years 2013 to 2018 – this will give an indication of publications published between one and six years previously (which parallels the time-window of the final contextual dataset).

Format

Contextual data will be provided to the REF team in two parallel formats in a Microsoft Excel spreadsheet: 1) as data tables; and 2) as a look-up table.

Data tables

For each of the three document types (i.e. articles, reviews and proceedings papers) we will provide a separate table containing contextual information. Each row in the table will contain data associated with a single Web of Science Journal Subject Category and data for each publication year will be shown in separate columns.

We will provide a separate sorted table of contextual data for each UOA, with subject categories corresponding to the greatest number of submitted outputs shown first.

We will also group the Journal Subject Categories into two less granular levels of classification to allow for easier filtering of some of the larger tables, to make identifying relevant subject categories more straightforward.

Two types of contextual information will be provided for each combination of document type, Journal Subject Category and publication year:

1. The **global mean number of citations per publication** in the relevant document type, category and publication year.
2. **Percentile thresholds**, i.e. the number of citations a paper would need to be ranked amongst the world's:
    - Top 1% highly cited papers
    - Top 5% highly cited papers
    - Top 10% highly cited papers
    - Top 25% papers by citation count
    - Top 50% papers by citation count.

While the global mean number of citations per publication provides an indicator of average performance, the distribution of citations with a group of publications is heavily skewed, with many papers that are less well-cited and relatively few highly cited ones. The percentile thresholds therefore provide an indication of how a paper's citation count compares to the global citation distribution for the relevant combination of document type, Journal Subject Category and publication year. These two types of contextual data taken together provide a more responsible basis for comparing the citation impact of research outputs.

Lookup

In additional to data tables containing all of the relevant contextual data, we will provide the REF team with a look-up worksheet in Microsoft Excel format to facilitate the retrieval of the same contextual citation indicators by document type, Journal Subject Category and publication year.

The goal of this additional worksheet is to increase the speed and flexibility of working with the contextual data.

The lookup spreadsheet will only be provided in the final snapshot of the contextual data provided at the close of the REF submission phase.

To use the look-up worksheet, users will manually select the document type, Journal Subject Category and publication year. The worksheet will return the above listed contextual data points for the relevant combination of these parameters.