Casting A Wide Net: The Journal Impact Factor Numerator

Stephen C. Hubbard; Marie E. McVeigh

Stephen C. Hubbard

Senior Editor, Journal Citation Reports

Marie E. McVeigh

Director, JCR and Bibliographic Policy

Thomson Reuters

1500 Spring Garden Street

Philadelphia, PA 19030

ts.production.tsagjcr@thomsonreuters.com

Phone: +1-215-823-3900

Abstract All metrics published in the Journal Citation Reports™ are dependent on the complete and correct aggregation of citations to each journal title. Here, we explain how unique cited titles are created for Thomson Reuters indexing, and how variations and ambiguities in title are collected in order to create the Journal Impact Factor numerator.

Introduction

The Journal Citation Reports (JCR) has been published annually since 1975. It offers a systematic, objective means to evaluate the world's leading journals in the natural and social sciences, with sortable journal- and category-level metrics derived from citation data. Although only journals in the Science Citation Index-Expanded and the Social Sciences Citation Index are listed in the JCR, the data underlying the metrics are drawn from all five Citation Indexes that comprise Web of Science: Science Citation Index-Expanded, Social Sciences Citation Index, Arts & Humanities Citation Index, Conference Proceedings Citation Index- Science, and Conference Proceedings Citation Index- Social Science & Humanities. In 2009, the coverage of these indexes includes nearly 12,000 active journal titles and over 3,000 published proceedings volumes, resulting in over 110,000 issues, 1.8 million indexed source items, and 48 million cited references that were analyzed during production of the JCR.

One of the key JCR metrics - certainly the most discussed - is the Journal Impact Factor. Much has been written on the Journal Impact Factor's strengths and limitations as a metric[1-5], and many alternatives have been proposed to compensate [6-10]. The Journal Impact Factor, in fact, has generated a large, and growing body of literature around its use and mis-use, its relevance and its irrelevance, its imminent demise, and its continued importance. Despite the complexity of reaction, the Journal Impact Factor is a simple ratio between the citations to a journal and the size of that journal. More specifically, the Journal Impact Factor is defined as the number of citations in the current JCR year to items published in the previous two years, divided by the total number of scholarly citable items published in those same two years. The 2009 Journal Impact Factor is calculated as:

2009 Journal Impact Factor = (citations in 2009 to items in 2008 + citations in 2009 to items in 2007) / (scholarly citable items in 2008 + scholarly items in 2007)

The numerator of this ratio considers the journal as a whole, and includes any citations to the journal title [11]; thus it depends on accurate and complete aggregation of citations to a journal title. The denominator considers the journal as a collection of items that are likely to influence the scholarly literature, by way of citation; thus only the scholarly or "citable" items in the journal are included. The composition of the denominator, which has been described previously,[12] is based on analysis of the content of the journal and the bibliographic parameters of its published source items. The remainder of the current paper will focus on the preparation of the Journal Impact Factor numerator. We will discuss the methods used to define cited journal titles, as well as how we identify, disambiguate and collect citations to each journal title.

Unique Cited Title

Preparation of JCR data begins as soon as a journal is accepted for coverage in Thomson Reuters citation databases. For each title, a unique 20-character abbreviation, a "preferred cited title" is created by JCR editors. A 20-character abbreviation is necessarily concise. A lengthier field would result in a greater number of variant forms for each cited work, which could result in these citations being missed in the preparation of total citation counts[13]. This abbreviated title will function both as the primary cited title in Web of Science, and as the title for display in the JCR. Each abbreviation is designed to be as full and clear a representation of the title as possible. For the majority of JCR journals, the 20-character limit is easily managed. For journals with very long titles, complete, unique, and unambiguous representation of the title in 20 characters can be a challenge. Long titles may require aggressive truncation, and some title words may not be represented in the abbreviated title. Using the journal ***Prostaglandins, Leukotrienes and Essential Fatty Acids*** as an example: common abbreviation of the title words would result in a title – PROSTAGLANDINS LEUKOT ESSENT FATTY ACIDS – that is more than twice as long as the 20-character standard field for Thomson Reuters. An effort to include all title words would result in the title: PROS LEUK ESS FAT AC, or a similar truncation. While complete, this presents a highly cryptic title that is unlikely to be used in cited references. Searching the 70,000 titles of scholarly journals in Ulrichs Web™ Global Serials Directory, and Thomson Reuters production system of over 250,000 titles of books, journals and proceedings retrieved 31 titles that begin with "prostaglandin", three titles that begin with “prostaglandin” followed by “leukotriene” (including the former title of this journal: ***Prostaglandins, Leukotrienes and Medicine***), but only one title where these two are followed by “essential.” Therefore, a unique and identifiable representation of the title can be created using only the first three words: PROSTAG LEUKOTR ESS.

A more unusual, and more complex situation arises in the case of two publications having identical or nearly identical titles, such as the journals ***Educational Psychology*** and ***Educational Psychologist***. While the publications are distinct, the way they are cited may not be. Using only the title as given results in an inherent ambiguity in our title list, since both titles would be abbreviated as EDUC PSYCHOL. Thus it is imperative that we create unique 20-character abbreviations as the collection point for citations to the individual journals. This is most commonly accomplished by appending the city or country of publication to the cited title abbreviation(s). The two titles given as an example were assigned the abbreviations EDUC PSYCHOL-UK and EDUC PSYCHOL-US, respectively, to establish two, distinguishable bibliographic entities, each of which can be cited. The process of creating unique abbreviations is applied even if only one of the titles is currently indexed by Thomson Reuters. When only one member of an ambiguous pair is covered, however, we try to avoid adding non-title elements (like city or country) to the abbreviation.

Citation Variants

The creation of an unique abbreviation only prepares a title to receive citations in Thomson Reuters products; it does not determine how those citations will appear in the literature. No matter how simple and unique a journal title, there will always be variations in how the journal is referenced. Journals have different policies and practices regarding the style of citation, and citing authors can further alter the title. Variability exists in the literature itself and must be considered in the process of citation collection. In the process of indexing cited references, Thomson Reuters uses both algorithmic and human-operator applied processes to decrease variability in the cited work. For example, we remove all non-essential words (the, an, and, from, or, etc.) and punctuation (&, -, :, ;, etc) from a cited work. We also apply a series of standard abbreviation and truncation algorithms both to shorten cited works so that more title elements will be represented in the 20-character cited work field. For example, the words “Biology”, “Biologic”, “Biological”, “Biologically” and the strings “Biol.”, and “Biologic.” are all shortened to BIOL in a cited work. Finally, our indexing process compares each reference to our entire database of source materials containing over 50 million items. If the cited work is recognized, at this early step, as an item that exists in our bibliographic database, the cited work is captured as the 20-character preferred cited title established by the JCR editors in accordance with policy. Without changing the variation that exists in the primary literature, we thus limit the number of variations that are entered in our cited reference index.

To resolve remaining variation in the incoming cited references, once the Thomson Reuters preferred cited abbreviation has been created, JCR editors build a series of dictionary entries associating each preferred title with a set of variant forms of the cited work. These entries are derived using direct observation and analysis of cited references. JCR editors have available more than 1 billion cited references in Web of Science, and to specialized reports on citations that are prepared for each JCR production cycle. Variants can include alternative abbreviated forms, alternate spellings, common misspellings, typographical errors, part numbers, numbered supplements, and others. Any observed variant that can be associated unambiguously with the covered title is included in the dictionary.

These dictionary entries have two purposes. First, they function in the presentation of cited references in Web of Science, driving variant citations towards the preferred abbreviation to simplify cited reference searching. Second, they identify title variants that will be included in the aggregation of citations for JCR metrics. These two functions are fulfilled separately and independently so that the cited reference index in Web of Science can present necessary additional information about a citation without preventing that citation from aggregating to the title in the JCR. For example, the journal ***Acta Psychiatrica Scandinavica*** has published over 430 supplements in its 50 year history. The fact that a supplement is being cited should be preserved in Web of Science to allow a user to locate the source more easily; nonetheless, a citation to the supplement should be included in JCR as a citation to the journal. Therefore the dictionary will aggregate citations to ACTA PSYCHIAT SCAND S into the total citation count for ACTA PSYCHIAT SCAND, but Web of Science will still display ACTA PSYCHIAT SCAND S as the cited work.

Journals that are cover-to-cover translations represent an additional special case in the JCR. Consider the title ***Journal of Evolutionary Biochemistry and Physiology*** (a translation of the Russian journal ***Zhurnal Evolyutsionnoi Biokhimii i Fizilogii***). Thomson Reuters indexes thecover-to-cover English-language translation of this journal, meaning the complete content of the title, both its scientific and its editorial items, are presented in full in both English and Russian versions. In order for the JCR to represent the impact of this content, citations either to the English title or to the Russian title must be combined towards the citation count for the journal. The preferred cited form of the title is given as J EVOL BIOCHEM PHYS+; the plus-sign indicates that citations to all English and Russian language variants have been collected for the JCR.

Currently, the Thomson Reuters production system includes over 210,000 variant title entries, for the resolution of citations to nearly 30,000 currrent and historic preferred titles. The average entry for a covered journal contains 10 variants, but approximately 1% of preferred titles require more than 50 variants.

Citation Aggregation for Journal Citation Reports

The curation of journal titles, creation of unique cited titles and collection of variant cited forms, comprise the backbone of citation unification at the journal level. “Casting a Wide Net” refers to the fact that the JCR is optimized to be as inclusive as possible in collecting citations to the journal as a whole, independently of whether the cited reference is linked to a specific source item in Web of Science. JCR will aggregate all citations to any recognizable variants of a journal’s title, and will distribute the citations according to the year of content referenced. For most journals, it is sufficient to use only the cited work and the cited year. The application of the dictionary of variants will allow a final aggregation of all citations based on the cited work, and the cited year will allow us to identify only those citations that refer to the prior one or two years of publication. For example, a cited reference in the form: "Abraham J (2008). *Astroparticle Phys.*" is clearly an acknowledgment of the journal ***Astroparticle Physics*** and should be aggregated to the JCR record for ASTROPART PHYS; however, there are two articles by J. Abraham in ***Astroparticle Physics*** in the year 2008. The incomplete reference data will not allow linking of this cited reference to either of the source items to which it could refer, but the JCR aggregation, working with just cited work and cited year, will include this citation in the 2009 Journal Impact Factor for ASTROPART PHYS. A cited reference that is too ambiguous to link to a source item can still be collected in the Journal Impact Factor numerator.

It is a different matter, however, when the cited reference is ambiguous as to the journal to which it refers, so let us revisit the journals ***Educational Psychology*** and ***Educational Psychologist***. Here, the need is not to decrease variation in the cited title, but to take one variant and divide citations accurately between titles. Most of the variants of the cited titles could refer to either title, such as *Ed. Psych.*, *Ed. Psychol.*, *Educ. Psych.*, *Education. Psych.*, *Educational Psychol.*, and others. A reference that reads "Chan, DW (2007). *Ed. Psychol.* **27**(1): 33-49." has a cited work ED PSYCHOL which could refer to either journal. For this kind of an ambiguous journal pair we use additional metadata elements in the cited reference to identify the specific journal referenced. It is important to note, however, that citation aggregation does not depend on whether the cited reference can be linked directly with a specific item in the journal. Linking a cited reference to a source item is a precise, point-to-point connection and will use either a few unique data elements, or several elements with increasing specificity; aggregating citations to the journal is more broad and will use a minimum of data to ensure accurate journal-level attribution. In this particular case it is sufficient to use the fact that ***Eduational Psychologist*** (EDUC PSYCHOL-US) published volume 42 in 2007, while ***Educational Psychology*** (EDUC PSYCHOL-UK) published volume 27 in 2007. Thus, the volume and year are sufficient to direct this citation to EDUC PSYCHOL-UK in the JCR.

In March of 2010, when we began production of the 2009 JCR data, the Thomson Reuters database contained over 53 million cited references that were indexed for the 2009 publication year. Forty-eight million of these references were published in journals and proceedings that are covered in Web of Science and so were extracted for analysis in the JCR. The first steps of unification apply abbreviation algorithms and unification of identical cited references, collapsing 48 million cited references to 6.4 million unique cited works. Analysis of ambiguous citations and application of the dictionary of variants results in a reduction to 1.4 million unique works, of which 12,212 are titles covered by Thomson Reuters.

Conclusion

The metrics in the JCR require simple mathematics, but this is just the final step in a continuous process of curation, analysis, editing and correction. Citation aggregation for each journal begins upon the addition of the title to our coverage list, and requires research and planning to create an unambiguous 20 character title as the principle identification of the journal. This preferred title is not just a representation of the journal for display. Rather, it functions as an essential part of the capture and aggregation of all citations to the journal. Variations in how a title is cited are added as dictionary entries so that citation to all unambiguous forms of the title can be collected to the unique 20-character preferred title. Ambiguous cited works are verified through additional metadata in the cited reference to allow clear attribution. Finally, citation counts are tabulated for each journal to construct the Journal Impact Factor numerator.

Through citation, scholars indicate work that has influenced them, reflecting upon, and synthesizing past knowledge. The metrics comprising the Journal Citation Reports™ collect these individual acknowledgements at the journal level, providing a way to identify those journals that consistently publish influential works. Production of the JCR continues throughout the year, in the continuous curation of titles indexed in Thomson Reuters citation indexes. This careful bibliographic control is driven by the goal of correct and complete identification of a cited work in each year of the JCR analysis.

1. GARFIELD E. Citation analysis as a tool in journal evaluation - Journals can be ranked by frquency and impact of citations for science policy studies. . Science 1972;178(4060):471-+.
2. SEGLEN PO. Why the impact factor of journals should not be used for evaluating research. Br Med J 1997 Feb;314(7079):498-502.
3. MONATSTERSKY R. The Number That's Devouring Science. Chronicle of Higher Education 2005 Oct 14;52(8):A12.
4. Not-so-deep impact. Nature 2005 Jun;435(7045):1003-4.
5. LIPPI G. The impact factor for evaluating scientists: the good, the bad and the ugly. Clinical Chemistry and Laboratory Medicine 2009 Dec;47(12):1585-6.
6. LEYDESDORFF L, OPTHOF T. Scopus's Source Normalized Impact per Paper (SNIP) Versus a Journal Impact Factor Based on Fractional Counting of Citations. Journal of the American Society for Information Science and Technology Nov;61(11):2365-9.
7. ZITT M. Citing-side normalization of journal impact: A robust variant of the Audience Factor. Journal of Informetrics Jul;4(3):392-406.
8. SOMBATSOMPOP N, MARKPIN T. Making an equality of ISI impact factors for different subject fields. Journal of the American Society for Information Science and Technology 2005 May;56(7):676-83.
9. BENSMAN SJ. The impact factor, total citations, and better citation mouse traps: A commentary. Journal of the American Society for Information Science and Technology 2007 Oct;58(12):1904-8.
10. ZITT M, SMALL H. Modifying the journal impact factor by fractional citation weighting: The audience factor. Journal of the American Society for Information Science and Technology 2008 Sep;59(11):1856-60.
11. BENSMAN SJ, LEYDESDORFF L. Definition and Identification of Journals as Bibliographic and Subject Entities: Librarianship Versus ISI Journal Citation Reports Methods and Their Effect on Citation Measures. Journal of the American Society for Information Science and Technology 2009 Jun;60(6):1097-117.
12. MCVEIGH ME, MANN SJ. The Journal Impact Factor Denominator Defining Citable (Counted) Items. Jama-Journal of the American Medical Association 2009 Sep;302(10):1107-9.
13. ROBERTSON J. Cited Title Unification. http://thomsonreuters.com/products_services/science/free/essays/cited_title_unification.