Web Indicators for Research Evaluation

A Practical Guide

Synthesis Lectures on Information Concepts, Retrieval, and Services

Gary Marchionini, *University of North Carolina, Chapel Hill*

Synthesis Lectures on Information Concepts, Retrieval, and Services publishes short books on topics pertaining to information science and applications of technology to information discovery, production, distribution, and management. Potential topics include: data models, indexing theory and algorithms, classification, information architecture, information economics, privacy and identity, scholarly communication, bibliometrics and webometrics, personal information management, human information behavior, digital libraries, archives and preservation, cultural informatics, information retrieval evaluation, data fusion, relevance feedback, recommendation systems, question answering, natural language processing for retrieval, text summarization, multimedia retrieval, multilingual retrieval, and exploratory search.

Web Indicators for Research Evaluation: A Practical Guide
Michael Thelwall

Trustworthy Policies for Distributed Repositories
Hao Xu, Mike Conway, Arcot Rajasekar, Jon Crabtree, Helen Tibbo, and Reagan W. Moore

The Notion of Relevance in Information Science: Everybody knows what relevance is. But, what is it really?
Tefko Saracevic

Dynamic Information Retrieval Modeling
Grace Hui Yang, Marc Sloan, and Jun Wang

Learning from Multiple Social Networks
Liqiang Nie, Xuemeng Song, and Tat-Seng Chua

Scholarly Collaboration on the Academic Social Web
Daqing He and Wei Jeng

Web Indicators for Research Evaluation: A Practical Guide
Michael Thelwall
www.morganclaypool.com

ISBN: 9781627059176 print
ISBN: 9781627059718 ebook

DOI 10.2200/S00733ED1V01Y201609ICR052

A Publication in the Morgan & Claypool Publishers series
SYNTHESIS LECTURES ON INFORMATION CONCEPTS, RETRIEVAL, AND SERVICES #52

Series Editor: Gary Marchionini, University of North Carolina, Chapel Hill

Series ISSN 1947-945X Print 1947-9468 Electronic

Web Indicators for Research Evaluation

A Practical Guide

Michael Thelwall
University of Wolverhampton

SYNTHESIS LECTURES ON INFORMATION CONCEPTS, RETRIEVAL, AND SERVICES #52

MORGAN & CLAYPOOL PUBLISHERS

ABSTRACT

In recent years there has been an increasing demand for research evaluation within universities and other research-based organisations. In parallel, there has been an increasing recognition that traditional citation-based indicators are not able to reflect the societal impacts of research and are slow to appear. This has led to the creation of new indicators for different types of research impact as well as timelier indicators, mainly derived from the Web. These indicators have been called altmetrics, webometrics or just web metrics. This book describes and evaluates a range of web indicators for aspects of societal or scholarly impact, discusses the theory and practice of using and evaluating web indicators for research assessment and outlines practical strategies for obtaining many web indicators. In addition to describing impact indicators for traditional scholarly outputs, such as journal articles and monographs, it also covers indicators for videos, datasets, software and other non-standard scholarly outputs. The book describes strategies to analyse web indicators for individual publications as well as to compare the impacts of groups of publications. The practical part of the book includes descriptions of how to use the free software Webometric Analyst to gather and analyse web data. This book is written for information science undergraduate and Master's students that are learning about alternative indicators or scientometrics as well as Ph.D. students and other researchers and practitioners using indicators to help assess research impact or to study scholarly communication.

KEYWORDS

web indicators, altmetrics, webometrics, alternative indicators, scientometrics, bibliometrics, scholarly communication, social media metrics

Contents

CHAPTER 1

Introduction

The need for research evaluation has increased dramatically in recent decades. A major source of demand is from organisations involved in the research process that need to assess the value of academic research (e.g., Wilsdon, 2016; www.snowballmetrics.com). These include:

- governments judging the effect of recent policy changes;
- governments and funding bodies interested in the value for money of their research spending;
- higher education funders dividing annual block grants between universities, such as through the UK Research Excellence Framework (REF);
- universities deciding how to share out budgets between departments;
- funding councils allocating project grants;
- departments choosing researchers to appoint and promote;
- libraries renewing journal subscriptions; and
- university managers, administrators, publishers or other stakeholders seeking to second guess or plan for the outcomes of any of the above.

This final reason is particularly important because any evaluation that has substantial financial implications triggers a need to plan in the affected organisations. This gives them a degree of control over, or early warning about, the outcomes. For example, probably all UK universities conducted mock REF exercises in the years before submitting to REF2014 (e.g., Gray, 2015; Owens, 2013) and had a dedicated team of REF administrators and academics in order to maximise their university's scores. Individual researchers also sometimes need to evaluate research to plan for future organisational assessments or for a variety of different purposes.

- To self-assess their progress
- To choose a venue in which to publish their work
- To select articles to read from a large collection that match their digital library search
- To select articles to read from a new issue of a journal or from recent additions to a digital repository

Scholars who investigate science itself may also need to assess bodies of scholarship. Even though many scholars produce intangible knowledge and understanding, most evaluations take the initial simplifying step of focusing exclusively or primarily on tangible outputs, such as articles, chapters and books.

Assessments that focus on academic outputs can be slow and complex. This is because scholars tend to produce work that is densely written, understandable to a small number of people, and can only be fully evaluated within the context of a large pool of similar outputs (e.g., for their level of novelty and relative strength of methods). Even experts with the background knowledge to understand individual publications need time to read them and may disagree on their merits. For these reasons it is difficult for non-experts to properly assess the work of scholars, for peers to evaluate the multiple outputs of large groups of researchers and for interested parties to sift through masses of published work in order to find the most important and relevant items. Thus, there is a seemingly impossible need to evaluate academic research *without reading it*.

Many in the past have resorted to citation counts to help with this task with the hypothesis that more cited works tend to be better. From this assumption, it follows that citation counts can be used as a convenient proxy for academic impact or quality. This belief is supported by the argument that scientists cite to acknowledge influential prior work so that highly cited papers are important for the progress of science (Merton, 1973). While this is broadly true in many areas of scholarship, each document's references are likely to be an incomplete and biased reflection of the influences on the research (MacRoberts and MacRoberts, 1989). Thus, citation counts can be misleading in individual cases even though in aggregate they may be reasonable indicators of academic impact in many fields (van Raan, 1998; Moed, 2005). Moreover, there is a substantial time lag between a scholar conducting research and their outputs having enough citations to estimate their citation impact.

A more fundamental weakness of citation counts is that governments and other funders rarely want to finance research for its own sake. Instead, scholarship is a means to an end, such as more effective higher education, enhanced international competitiveness, improved public health and national cultural enrichment. In this context, citation counts appear to be measuring the wrong research outcome. They have been adopted in the past due to a lack of alternatives and arguments by academics that research excellence would itself naturally lead to the desired societal benefits. Nevertheless, there is ongoing pressure for indicators that will more directly reflect valuable research application types.

In parallel with the need to evaluate the non-academic impacts of research is the need to evaluate non-standard research outputs. Academia is increasingly complex and digitised (Meyer and Schroeder, 2015), with individual scholars and groups producing outputs of types that are often ignored in research assessments. For example, while evaluations may focus on academic journal articles and books, scholars may also produce software, databases, videos, blogs and other artefacts.

These can be central to a research field, as in the case of biodiversity and chemical databases, machine learning software environments and natural language processing toolkits. They can also make helpful contributions to education, as in the case of videos exploring the normally unseen work of scientists (www.test-tube.org.uk). Non-standard outputs can also support science as a whole as in the case of the blog posts and newspaper articles of Professor Stephen Curry (www.theguardian.com/profile/stephen-curry), which promote and explain issues of general importance to researchers. There is little hope that citation counts can be useful in evaluating many of these contributions, although there are drives to encourage researchers to formally cite the software and data that they use in their studies.

A similar issue is that academics are increasingly called upon to directly engage with potential end users of their research and to seek ways in which their expertise can help society. As an example, the Flanders Marine Institute assesses local marine biodiversity, has a school outreach programme and produces publications aimed at the local fishing industry, including *Compendium Coast and Sea*, which aims to "aggregate objective and scientifically-underpinned information and data from Flemish/Belgian marine and maritime research" (www.vliz.be/en/compendium-coast-and-sea). Similarly, the Oxford Internet Institute produces a periodic survey of internet use in the UK as a service for the community (oxis.oii.ox.ac.uk) and the *Stern Review on the Economics of Climate Change* is a high-profile example of a huge amount of academic work used to directly inform government policy. Any indicators that could help to assess the impact of such activities would be valuable for the funders that need to check that they are getting value for money.

A potential solution to the above problems has appeared in the form of the web. The rise of the web to become embedded in the work of scholars and wider society has created a situation in which there is easily accessible public online evidence of research impacts that could be exploited for new indicators. While none of the new indicators can deal with all research evaluation needs, some can reflect important non-academic types of impact or can be applied to non-standard outputs. Questions such as the following have triggered a particular interest in web indicators.

- Do tweet counts reflect the degree of public interest in research?
- Do citations from online patents reflect commercial technology transfer?
- Can mentions in the online grey literature provide evidence of policy impacts?
- Could the fast publishing nature of the web make it possible to generate early impact indicators?

There are strong opponents of the use of impact indicators and particular types of web indicators (Colquhoun and Plested, 2014) and some have argued that their value has been exaggerated (Barnes, 2015). Thus, it is important to critically evaluate the value of web indicators, to develop effective methods to use them and theory to help interpret them.

This book describes and evaluates a range of web indicators, drawing upon relevant empirical studies. It takes a critical perspective, emphasising the generic and specific limitations of each indicator so that they can be used with appropriate care. It also takes a positive approach by describing contexts in which web indicators can be useful and providing practical help with calculating them.

1.1 INDICATOR TERMINOLOGY AND INTERPRETATION

Indicator terminology is used somewhat interchangeably in practice, but it is useful to give precise definitions for this book. Here, an **indicator** is a number that is used or developed to point to the direction or level of achievement of an entity of any sort. It may be a simple individual number, such as a tweet count, or the result of a mathematical formula applied to a set of numbers, such as the arithmetic mean or geometric mean. This definition excludes non-numerical entities, such as pictures and network diagrams, although indicators can be represented by graphs or within network diagrams. In society, a well-known indicator is the gross domestic product, which provides evidence of national economic strength. In the UK, the retail price index (RPI) calculated by the Office for National Statistics (ONS) estimates the change in purchasing power of money for consumers by monitoring the average price of a sample of retail goods and services. Its inflation values are widely used despite being misleading for consumers who purchase goods or services that are not on the list and that have a different pattern of price changes, such as luxury or black market goods. It can also be misleading overall if an essential good or service that is not on the list exhibits a large price change that is out of line with other goods and services due to a market collapse or flooding. Despite the demonstrable failings of the RPI it is still a useful indicator of the purchasing power of money for typical consumers. This leads to the following conclusion.

> An indicator does not have to be very accurate to be useful as long as, on average, higher values associate with higher levels of the quantity being assessed.

This point is important because a common criticism of citation counts (and, by extension, most web indicators) is that research can be cited for negative reasons, such as to criticise methods or findings (MacRoberts and MacRoberts, 1996). Despite this, if indicators tend to give scores that agree to a large extent with human judgements then it would be reasonable to replace human judgements with them when a decision is not important enough to justify the time necessary for experts to read the articles in question.

> Indicators can be useful when the value of an assessment is not great enough to justify the time needed by experts to make human judgements.

For decisions that are important enough to require expert judgements on a collection of outputs, indicators can still be useful to cross-check the expert opinions in order to either highlight areas in which they may have overlooked excellence or overrated poor research—although

the experts should make the final decision. Indicators may also be used to help look for evidence of systematic biases in human judgements, such as on the basis of author gender, disability or ethnicity. For example, if the outputs of female academics tended to have indicator scores that were double those of male academics before being judged excellent then this should trigger gender bias investigations.

Indicators can be useful to support or cross-check expert human judgements.

Indicators can support specific impact claims by academics for individual outputs. For example, someone might argue that devoting all of their research time to a blog is valuable because of the large audience that it attracts. They could use blog access statistics to support this claim.

Indicators can support the impact claims of practicing researchers.

Indicators can also be preferable to human judgements for theoretical analyses of science itself and of systemic biases, such as in terms of the career success of researchers by gender, disability or ethnicity.

Indicators can be preferable to expert human judgements for system-level analyses.

Although perfect accuracy is not essential for any of these tasks, increased accuracy is clearly desirable for any indicator and increases the range of tasks for which it is useful.

1.2 METRICS AND INDICATORS

It is useful to distinguish between the terms *metric* and *indicator*, even though the difference between them is one of perspective and there are different uses of both terms in society and academia (e.g., Lazarsfeld, 1958; Hubbard, 2014). Both *metric* and *measure* can cause confusion in a scientometric context because a commonsense interpretation is that if something has been *measured* then the measurement will be essentially exact (e.g., Hubbard, 2014, p. 30). In science (including the social sciences), however, a measurement usually has a more technical interpretation as a quantitative entity that reduces uncertainty to any degree at all, and can incorporate even large errors (Hubbard, 2014, p. 32). Thus, for example, if it is possible to show that knowing the number of web citations received by an article reduces the degree of uncertainty about whether it is a good quality article or not then it would be scientifically reasonable to describe web citation counts as a *measure* of (research) quality. Nevertheless, it would be unreasonable from a commonsense perspective to describe web citation counts as a measure of quality because highly cited articles can be poor and so the web citation count measurement can have very large errors. The same logic also holds for traditional citation counts and all web indicators. Thus, the terms *measurement* and *metric* carry commonsense connotations of accuracy, even though this is not a scientific property of their common definitions.

For research evaluation purposes, although not common practice, it would be helpful to reserve the terms *metric* and *measurement* for things that have a reasonable degree of accuracy. This

would reduce the frequent commonsense objections to indicators on the basis of their large obvious errors. These objections often take the form of individual cases of large discrepancies, such as highly cited false papers, or evidence of bias. For example, the retweet count reported by Twitter is a retweet metric for the number of times that a tweet has been retweeted. Presumably this number tends to be reasonably accurate even though it may occasionally be wrong if very recent retweets are ignored by the software that calculates it. In contrast, it could be misleading (although not technically incorrect) to describe the retweet count as a popularity *metric* in Twitter because a tweet may be highly retweeted by spammers and so retweet counts can give highly inaccurate popularity estimates in some cases.

It is also helpful in research evaluations to avoid using terminology with connotations of accuracy because the outcomes may be used by non-experts to judge researchers and these non-experts may apply a commonsense interpretation. This may lead to unwarranted accuracy assumptions (e.g., that all articles in journals with high impact factors are excellent) or the rejection of all data because of occasional obvious large errors. Both of these over-interpretations are reasonable from a commonsense understanding of the term *measure*.

In contrast, the term *indicator* does not carry the strong commonsense connotations of precision that the terms measure and metric do. Although (again) not common practice, it is helpful in research evaluation to use *indicator* for any quantitative entity that is known or believed to associate with the phenomenon of interest even if it does not accurately measure it. In the above case retweet counts could be described as a popularity *indicator* within Twitter because it seems reasonable to believe that, in general, more popular tweets will be retweeted more than less popular tweets. Confusingly, an indicator can be a (commonsense) metric for one thing but not another. The counts of tweets citing an article is a metric of how often the article is tweeted, and may be an indicator of the impact of the article but should not be described as an impact metric for articles because counting tweets does not accurately measure the impact of an article. In this book, the term *metric* will be used for indicators when discussing that they measure something (other than research impact) with a reasonable degree of accuracy.

> For research evaluation it is better to use the term *indicator* than the terms *metric* or *measurement* for quantities that can have large errors. This will help to reduce confusion from non-experts who may assume that metrics and measurements should be accurate.

The definition of an indicator does not include any cause-and-effect requirement with the type of impact assessed. While showing that such a relationship is present would help to interpret the meaning of a web indicator, it is not necessary. If there is an unknown, weak or no cause-and-effect relationship then extra caution should be used when interpreting indicator values, however.

In practice, this means that an indicator should not be used as the primary source of evidence, if possible, and that its values should be checked for anomalies in each application.

Additional problems arise when indicators are used to help assess researchers, who may focus on the indicator rather than the goal of the assessment. In a policy context, whenever an indicator is publicly selected in advance of an evaluation then getting higher indicator values is desirable for those evaluated or other stakeholders. If there is not a straightforward and robust cause-and-effect relationship then there will be a behaviour change in the direction of the indicator rather than in the direction of the factor that it is an indicator of (e.g., getting friends to retweet rather than trying to write more popular outputs). Thus, consideration of the consequences of using indicators is needed as well as information about their properties.

> The consequence of the use of an indicator on the behaviour of those assessed must be thought through before the indicator is used.

A partial exception to the above is for "surprise" evaluations in which the indicators to be used are not known in advance but are requested by the judges performing the assessment. In such cases, it is too late for those assessed to modify their behaviour but there may still be behaviour change after the assessment if they anticipate the use of similar indicators in a future assessment.

1.3 WEB INDICATORS

This book focuses on academic and academic-related indicators derived from the web. Citation counts from the Web of Science and Scopus are the most researched type of academic indicator and are the benchmark for discussing alternative indicators. Here, a **web indicator** is a number that is (a) intended to associate with an aspect of research performance or impact, and that is (b) derived from the web and not in any way from counts of citations from academic journal articles. This excludes citation counts from the Web of Science and Scopus (even though they are on the web) as well as formulae that process citation counts, such as the Journal Impact Factor (Garfield, 1999) and field-normalised citation counts (Waltman et al., 2011). In practice, within this book, web indicators always relate to tangible academic-related outputs, such as journal articles, scholar-produced videos, or monographs. They can reflect general, academic, educational, commercial, organisational or information impact.

There are many different types of web indicators. The most well-known type, called social media metrics or altmetrics (Priem et al., 2011; see also: Holmberg, 2015), are derived from social websites, such as Twitter, that are free to join and open to the public. Social media metrics are typically collected by a computer program through an applications programming interface (API), and this facility has made them relatively easy to collect. A more general term is *webometrics* (Almind and Ingwersen, 1997), which originally referred to all indicators derived from the public web and now also describes a research field of the same name. The word *webometrics* is currently used for

indicators derived from the web except for social media metrics. Usage metrics, in contrast, give evidence of how often a document or other resource has been viewed, downloaded or otherwise accessed online (Kurtz and Bollen, 2010). Usage metrics may be derived from the web or social web, overlapping with the previous two classes, or may be derived from web server log files, which are about the web but are not themselves part of the public web or social web. These are included in this book for completeness.

This book does not cover alternative indicators that are derived primarily from non-web sources, such as in the case of some patent and reputation indicators.

1.4 BOOK-SPECIFIC INDICATORS

Although most research into web indicators has focused on journal articles, in the arts, humanities and some social sciences, important scholarly output types include monographs, edited books and book chapters. While it is possible to count citations to books from journal articles in traditional citation indexes, such citations do not reflect the scholarly and other impacts of books well (Cronin et al., 1997). This is because books can target book-based research fields. The situation has been partly remedied by the Web of Science and Scopus indexing substantial numbers of books, but both cover only a small, English-focused subset of the world's academic output and there are better web indicator solutions.

Books in general can productively be used in many ways that do not lead to new journal citations, such as supporting education, informing policy, professional practice and health behaviours, and culturally enriching the reader. Books may also further lines of research that are predominantly published in monographs and book chapters. In addition, citation practices are different in the humanities, with serendipitous citations being common (Stone, 1982) even though they do not reflect the academic contributions of the cited works. Nevertheless, it seems reasonable to believe that important books would tend to be highly cited, whether by other books or by journal articles. The best source of data about citations to books is therefore a huge book database, and Google Books is the logical choice for this. Web indicators can be particularly helpful for books because of the multidimensional ways in which they can have impacts (Halevi et al., 2016).

It is particularly useful to have evidence of the number of readers for a book since books do not need to generate citations in order to have an impact, but they must at least be read first. The ideal book readership evidence would be book sales (print and electronic) added to library loans but sales information does not seem to be released by publishers and may not be reliable between publishers. Similarly, library loan information does not seem to be ever put in the public domain. Proxy readership indicators are therefore needed. Mendeley reader counts would be a logical choice but users seem to rarely register books on the site. One public source of sales-related data, albeit only from specific online bookstores, is the sales rank published by sites like Amazon.com. For

libraries, although lending information is not shared, except perhaps with the security agencies, catalog information is usually public and so it is possible to count the library holdings of a book as a proxy indicator of the likely extent of its readership.

1.5 INDICATORS FOR NON-STANDARD SCHOLARLY OUTPUTS

Data, software, videos, blogs, reports and images are important outputs of some scholars' research, and web indicators, such as view or download counts, are a natural source of impact evidence for them. Non-standard scholarly outputs that give value to the scholarly community need to be recognised so that researchers continue to create them and do not divert their attentions to less valuable but more recognised activities. This book briefly discusses indicators for a range of different types of academic output. The brief coverage is partly due to the scarcity of relevant investigations and partly because, as discussed in the conclusions, web indicators for non-standard outputs are best used in a simple way.

The magnitude of any indicator should be interpreted relative to the context in which it is expected to be used and this can vary enormously for non-standard scholarly outputs. In theory, the impact of a resource should be the number of uses times the average value of each use. On this basis, for example, an icon library with millions of fairly trivial uses could be fairly compared against a video illustrating a new separation technique for conjoined twins that may only have a few viewers but each one might save two lives. Such calculations are impossible because the users of a resource are often unknown and the value of their uses would typically be impossible or impractical to calculate. Of course, the same is true for citations: while some are fairly trivial, others can be vital to new studies. Nevertheless, the magnitude of the difference is much larger for most resource types, except perhaps datasets because it is difficult to imagine that there are many trivial uses of datasets. Because of this, it seems most useful to employ web indicators to individual resources or homogenous collections of resources (e.g., Haran and Poliakoff, 2011) accompanied by a textual explanation of the context. The case for the value of an indicator could be strengthened if it could be benchmarked against equivalent values for resources of a similar type and intended audience. This would allow resource owners to make claims such as "X is the most downloaded free software for counting angels on pinheads" or "Our video of fly eggs hatching in horse manure has been viewed twice as often as video of the same event that is also aimed at school pupils." Unfortunately, benchmarking is difficult in practice because the well-known resources tend to be the successful minority while the failures are unknown and difficult to find.

1.6 DISCIPLINARY AND TIME DIFFERENCES

The typical values of all indicator data vary by field and year (see Section 9.3). Because of this, indicators should not be calculated for sets of articles from multiple fields and years and should not be compared between fields and years. More sophisticated indicators are needed to compare between fields and years and these are discussed in Chapter 9.

> Most indicators should not be compared between fields because of disciplinary differences. Most indicators should not be compared between years because of time differences.

1.7 OVERVIEW AND INTENDED AUDIENCE

This book describes a range of web indicators for different impact types, covers the theory and practice of using web indicators for research assessment and outlines practical strategies for obtaining many web indicators. It discusses the use of indicators to help evaluate traditional scholarly outputs, such as journal articles and monographs, as well as for other types of online scholarly outputs, such as videos, datasets and software. The book also describes how to evaluate collections of such outputs, such as those produced by individuals, groups and institutions.

This book is aimed at undergraduate and Master's degree students within information science who are learning about alternative indicators or scientometrics, as well as Ph.D. students, researchers and practitioners who are using, or would like to use, alternative indicators for academic impact evaluations or to study scholarly communication. The first part is also aimed at policy makers and research administrators who need to know which indicators to use and how to interpret their values. In contrast to a volume on webometric methods for investigating sets of websites (Thelwall, 2009), and recent excellent and thought-provoking books about altmetrics and alternative indicators (Holmberg, 2015; Cronin and Sugimoto, 2014), the focus of this book is on bringing together in one place the evidence, methods and tools needed to use web indicators in research evaluations.

As part of the goal to provide practical help, this book includes detailed instructions about how to use the free software Webometric Analyst to calculate simple and advanced indicators. This information is supplemented by the website hosting the software, which gives updates and additional help. The book does not attempt to be comprehensive in this regard and so does not explain how to process the free data provided by Altmetric.com to researchers, or how to calculate every possible indictor formula.

Chapters 1 to 7 and 11 target all readers, whereas the more technical Chapters 8 to 10 covering statistical and software issues are mainly designed for people who wish to collect their own web data and calculate a range of indicators from it. Chapter 9 on statistics is also useful for policy makers and research administrators. On a stylistic note, in places this book contains dense lists of

correlation coefficients from different studies. These are usually restricted to paragraphs labelled "empirical evidence" and can be skipped on a first reading.

CHAPTER 2

Evaluating Indicators

Indicators must be evaluated before they can be used with any confidence. Evaluations can assess the type of impact represented by the indicator and the strength of the evidence that it provides. For example, while it may seem obvious that the public uses Twitter, and so counts of tweets about articles would be a useful indicator of public interest, this is usually not true (e.g., Thelwall et al., 2013b; but with some exceptions: Desai et al. 2012), and so tweet counts should not be used as indicators of public interest. Even the very general assumption that articles that are mentioned often in the social web tend to be more important needs supporting evidence. Web indicators also need to be evaluated because articles may be mentioned on the social web for negative reasons, such as to criticise them (Shema et al., 2012), to accuse the authors of fraud, to discuss retracted papers (Marcus and Oransky, 2011), for irrelevant reasons such as spam, or automated mentions (e.g., a journal tweeting all its articles, when published) or because they have funny or interesting titles. If alternative indicators are to be taken seriously in evaluations, then concrete evidence is needed to justify their use to those evaluated.

Some web indicators have an obvious face value interpretation, but these still need to be evaluated for evidence of bias and prevalence. An example is the online syllabus mention indicator. Unlike the case of tweet citations mentioned above, a citation from a course syllabus can be taken at face value as evidence of educational impact because course syllabi are created by instructors and the cited works contained in them are intended to be read by students as part of their education. Nevertheless, it is not immediately clear whether it would be fair to compare the educational impact of articles based on their online syllabus mentions because research can be used in education without being cited if it is summarised in standard course textbooks, part of a field that rarely recommends readings to students or included in syllabi that are not placed online. In addition, if very few course syllabi are posted to the public web then syllabus mention indictors would have too low coverage of academic research to have much practical value. A more general point is that a score of 0 on any particular indicator does not imply that the output assessed has had no impact, but only that no impact was recorded for it by the indicator.

For all of the above reasons, it is important to validate alternative indicators before use. Evaluations of indicators are not simple, however. Even citations, which are produced in a quality controlled environment (i.e., scholarly peer reviewed journals) and have been researched for decades, are controversial in two senses: whether they should be used at all (MacRoberts and MacRoberts, 1996; Seglen, 1998) and how their meaning should be interpreted (i.e., what they indicate) (Moed,

2005). The rest of this chapter discusses a range of accepted evaluation methods for indicators and makes overall recommendations for evaluation strategies.

The validation process is different from a common social sciences model of starting with a concept and then producing a series of measurements to capture it as well as possible (e.g., Lazarsfeld, 1958). The impact indicators discussed in this book (e.g., tweet citation counts) are instead essentially available before the concept rather than constructed to measure the concept. In consequence, the validation process needs to assess whether the indicator broadly reflects the type of impact that it appears to, as well as to find out in more detail about the type of impact, if any, that it represents.

A complicating issue is that it may be difficult to extract comprehensive data before calculating an indicator. The lack of a complete directory of blogs, for example, means that it is impossible to count all citations from blogs. It can also be difficult to extract accurate data. Counting the number of people who bookmarked an article online, for instance, may be difficult if some people maintain multiple social web bookmarking accounts and others share accounts. Indicators may also be systematically biased by marketing initiatives, such as authors, journals or institutions tweeting all of their articles. Most significantly, if an indicator becomes highly valued then authors, editors, or publishers may attempt to artificially inflate their scores. The lack of a quality control mechanism within the web makes deliberate and accidental manipulation difficult to stop.

2.1 EVALUATION METHODS

This section evaluates indicator data (e.g., tweet counts) rather than any specific indicator formula (e.g., median tweet counts). A range of methods has previously been used to investigate academic-related indicators. Correlation tests are the most common, but are insufficient on their own and particularly for web indicators that claim to reflect something other than scholarly impact.

2.1.1 HOW COMMON? COVERAGE ASSESSMENT

The prevalence of an indicator affects its usefulness; if a tiny fraction of articles received a non-zero count for a given indicator then it would have little value for applications that rely upon scores for individual articles, such as article altmetrics in a digital library. Nevertheless, low coverage does not preclude all applications because indicators with low coverage can still be used to compare the impact of groups of documents. This is possible by comparing the proportion that have a non-zero score between groups. For example, if group A had a higher proportion of articles with a non-zero score on a particular educational impact web indicator than group B then this would give some evidence of higher overall educational impact from group A. In general, the lower the coverage of the indicator, the larger the groups of articles that would need to be compared to detect differences between them.

2.1.2 CORRELATIONS WITH PEER REVIEW OR CITATION COUNTS

The most practical technique to help validate a research indicator is to calculate the correlation between it and a better understood data source, such as citation counts or peer review scores, even though these have their own biases (Lee et al., 2013; MacRoberts and MacRoberts, 1996; Wennerås and Wold, 1997). Correlations have been extensively used in webometrics to evaluate the evidence provided by links to journal websites or individual articles (Vaughan and Huysen, 2002; Vaughan and Shaw, 2003, 2005) or URL citations (Kousha and Thelwall, 2007) to articles or citations from various parts of the web (e.g., Thelwall and Kousha, 2008). They have also been introduced for altmetrics, playing a similar role (Li et al., 2012). Spearman correlations are normally used because citation data is typically too skewed for the normality assumption of a Pearson test. Given the relative trustworthiness of peer judgements, the best correlation would be between a rank order or scores produced by peer review and the rank order produced by the indicator. In practice it is difficult to get appropriate experts to rate lists of publications and so citation counts are routinely used instead on the basis that citations are an established research impact data source.

The rationale for calculating the correlation between an indicator and another source of research evidence (e.g., peer review rankings or citation scores) is that if they both reflect a type of research impact then the two rankings should be related, giving rise to a positive correlation coefficient, even if they reflect different types of research impact. In the hypothetical case that two indicators both measure pure research quality (assuming that this exists) then their correlation would always be positive, with a magnitude determined only by the amount of natural random fluctuations in the data. In the more realistic case that both partly reflect different aspects of research impact (e.g., educational utility or value for future scholarship) then the extent of the correlation would also depend upon how closely related these two aspects were. Finally, most metrics also probably reflect unwanted systematic causes of bias (e.g., institutional bias or time-dependency) which also affects the magnitude of a correlation and may even change its sign.

For web indicators, a positive correlation with citation counts gives evidence that the indicator at least partly reflects academic quality. This is because citation counts are known to partly reflect academic quality to some extent in most fields and so should correlate positively with any other indicator that also correlates with research quality. There is a gap in this logic because it is possible that citation counts and a web indicator have a positive correlation because they both reflect the same aspect of articles that is irrelevant to research quality, such as the publication language. Another gap is that the alternative indicator could have a zero correlation with citation counts because they both exclusively reflect completely different aspects of research quality. Nevertheless, these scenarios seem unlikely to occur in a pure form and so correlations with citation counts are an established test of association that is necessary to validate alternative indicators. A statistically significant positive correlation also gives evidence that the web indicator is not purely

random. Thus, the positive correlation with citation counts in Figure 2.1 gives some evidence that Mendeley readers are valid research quality indicators for orthodontics articles and strong evidence that Mendeley readership data is not purely random.

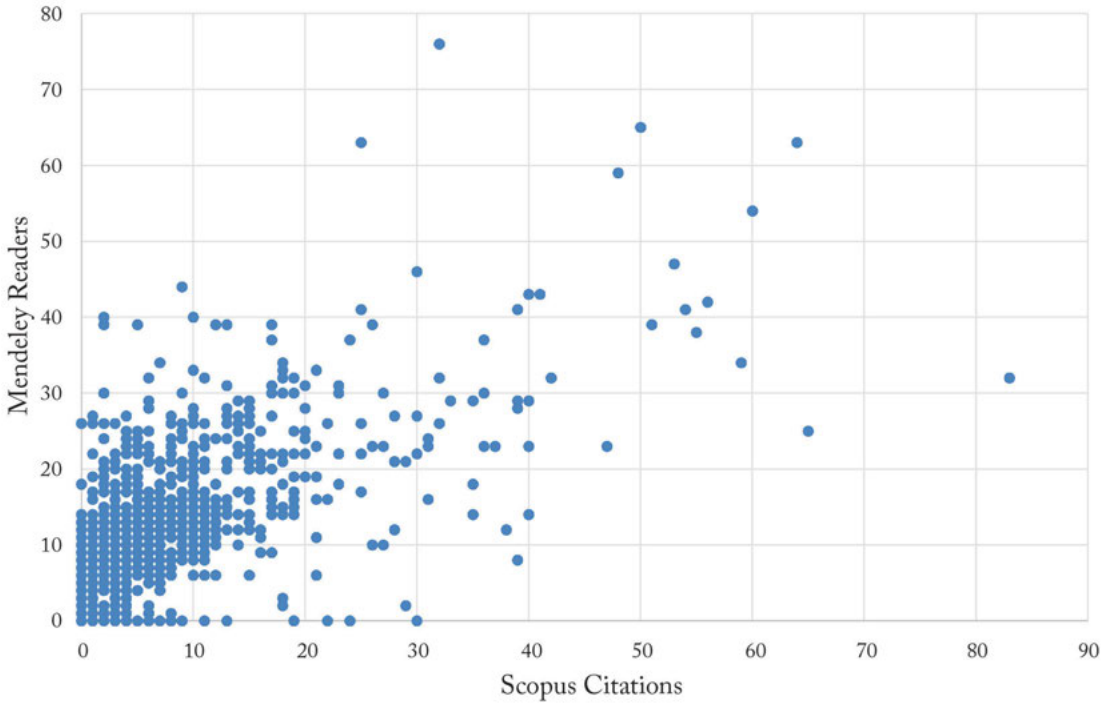

Figure 2.1: A scatter plot of Mendeley readers against Scopus citations for orthodontics articles from 2009. The graph shows a clear tendency for articles with many citations to also have many readers. A Spearman correlation of 0.744 reflects this strong relationship.

The potential for systematic causes of bias means that a positive correlation between a new and an established research indicator does not prove that the new indicator reflects an aspect of research quality because the correlation could be spurious and caused by a factor unrelated to research. Conversely, a negative correlation does not disprove the relationship because there may be an underlying research-related positive relationship that is suppressed by a factor unrelated to research. Hence, the onus is on the researcher to remove potential sources of bias as far as possible. For example, studies should focus on articles published within a limited time window to reduce the impact of time differences on the results. Collections of articles should also be as homogeneous as possible, such as by taking them all from the same journal or field and excluding reviews. In practice, positive correlations with citation counts are accepted as evidence of their value as indicators of an

aspect of research impact if there is no obvious source of bias in the comparison made. The normal requirement for the test is that the correlation coefficient is statistically significant and greater than zero. Correlation calculations in scientometrics seem to always use several hundred articles or more, which is an adequate sample size.

The magnitude of a correlation coefficient is important, with higher values presenting stronger evidence of the value of the indicator as well as evidence that the relationship between the indicator and citation counts is closer. The strength of a correlation coefficient can be greatly reduced if the sets of articles being compared are from different fields or years (Thelwall, 2016d) and so articles should be separated out by field and year before calculating correlations. Low numbers and many zeros can also reduce correlation values, masking the strength of the underlying relationship (Thelwall, 2016d). Alternative indicators that increase rapidly within an individual year may need to be evaluated with the sign test (Section 9.10) rather than a correlation test because of this.

The main limitation of correlation tests between alternative indicators and citation counts is that they cannot give evidence of the *type* of impact reflected by the web indicator, if it is different from citation impact. Thus, if a new web indicator for educational impact is proposed then a positive correlation with citation counts demonstrates that it is not random and relates to scholarly activities in some way, but does not show whether it gives evidence of educational impact. Correlation tests alone are therefore insufficient for indicators that claim to reflect any non-scholarly type of impact.

2.1.3 WHY? CREATOR MOTIVATION INTERVIEWS OR QUESTIONNAIRES

The most direct way to assess whether an indicator reflects a particular type of impact is to interview the creators of the raw data (e.g., the tweeters for the tweet count altmetric) to find out why they created the data (e.g., a tweet). If the reasons tended to at least partially align with a particular type of impact, then this would support the validity of the indicator for that type of impact. For example, if most tweeters interviewed claimed to only tweet links to articles that they considered to be *useful* for research then it would be reasonable to claim tweet link counts as research utility indicators. In contrast, if most tweeters reported different motivations, such as tweeting articles with funny titles, then tweet counts could not be claimed to be research utility indicators.

In practice it is likely that a range of motivations would be elicited by interviews (Priem and Costello, 2010) and so in order for an indicator to be useful then the dominant reason(s) should relate to a specific type of impact (e.g., educational, commercial) and the other reasons should not introduce systematic sources of bias (i.e., common biases), unless they are too rare to be significant.

Creator motivation interviews have featured in few studies for three reasons: they are time consuming; they can only include a few relevant web authors; and authors may not be reliable because they have forgotten, do not understand or mask their reasons (as is the case for citations:

Brooks, 1986; Case and Higgins, 2000). Nevertheless, such interviews may give insights that are known only to the creators of the data and would not be evident from other methods. For example, interviews with tweeting academics revealed that some tweeted on the basis of reading blogs discussing articles rather than the articles themselves (Priem and Costello, 2010). The scope for future qualitative research of this nature seems limitless because of the range of indicators and likely differences in uptake and styles of use between researchers based upon countries, disciplines, fields and ages. What would be particularly useful in this regard, therefore, would be theories that would help to generalise patterns of use so that the inevitable large gaps in knowledge (e.g., for unexamined countries or disciplines) would not cause problems. A corollary to this is that contextual information about the value of indicators from creator interviews or questionnaires is likely to always be patchy in terms of the fields and indicators investigated.

Creator motivation questionnaires are also rare partly because it is usually difficult to get a representative sample of creators. One exception used an elaborate process to get the email addresses of a large sample of Mendeley users (Mohammadi et al., 2016) in order to send them a questionnaire asking how and why they employed the service.

2.1.4 WHY? SOURCE CONTENT ANALYSIS

A practical alternative to author interviews or questionnaires is to conduct a content analysis of a random sample of raw data (e.g., tweets with citations) to categorise the contexts or the apparent citation motivations (Priem and Costello, 2010). This is non-intrusive, can be conducted on a larger scale than interviews or questionnaires, and does not rely upon author memories. Its disadvantages are that insufficient context may be available for a reliable classification in some cases, coders may be fooled by clever spam and it is labour-intensive to do well. The amount of context and hence the usefulness of this approach varies by data source. While tweets may be too short, blog posts should typically give enough context for reliable coding. Any content analysis should follow standard guidelines: using careful descriptions and multiple coders and reporting inter-coder reliability (Neuendorf, 2002).

Content analyses have been rarely used for alternative indicators (exceptions: Priem and Costello, 2010; Thelwall et al., 2013b) but deserve to be more common. In addition to giving evidence about why the raw data was created, which is essential to validate the type of impact reflected by indicators, they can improve the wider understanding of their meaning by revealing their typical contexts.

As for interviews, in order for a content analysis to provide evidence of research value in the associated indicator, the dominant (not necessarily the majority) category should be related to the type of impact claimed and the remainder should not introduce systematic sources of bias, unless they are much smaller. These provisos greatly complicate the interpretation of the results: unless

reasons related to a single impact type are in an overwhelming majority, a qualitative argument must be made for the remaining categories not introducing systematic biases.

2.1.5 WHO? CREATOR TYPE QUESTIONNAIRES AND DATA

In addition to finding out *why* indicator values are created, it is also important to know who creates them. For example, it would be useful to know who uses the social web for scholarly purposes and which parts they use (Weller et al., 2010; Procter et al., 2010). This information can point to systematic biases, such as towards younger users or females. This can be investigated using questionnaires, with the same considerations as above.

The demographics of the creators of indicator data can also sometimes be investigated directly from the source by extracting information about them from the web. In one study, properties of Mendeley readers (e.g., academic status and nationality) were harvested from the Mendeley.com API in order to give large-scale information (Mohammadi et al., 2015).

2.1.6 PRAGMATIC EVALUATIONS

A final type of evaluation is pragmatic (Helic et al., 2011): testing whether a specific use of an indicator helps to achieve a desired goal. In other words, this means evaluating the use of the indicator in practice. In research assessments a pragmatic evaluation would involve discovering the opinions of some or all of the participants about their perceptions of the usefulness of the indicators provided. Depending upon the scale of the evaluations, this could take the form of interviews or questionnaires. Assessors could be asked whether they felt that the indicators helped them to arrive at a more accurate or quicker judgement. Such assessments were conducted informally for the UK REF2014 assessment to discover which disciplines found citation counts to be useful indicators. Similar sessions do not seem to have been conducted yet for any alternative indicator. For indicators displayed on a publisher's website, a pragmatic evaluation might instead ask users whether they believed that altmetrics helped them to find important or useful articles.

2.2 DELIBERATE AND ACCIDENTAL MANIPULATION OF RESULTS

A problem that affects typical web indicators is that they can be manipulated due to a lack of quality control. Accidental manipulation might occur, for example, though publicity for articles by their authors and publishing journals. Assuming that accidental manipulation of this type is ongoing at a constant level, the techniques discussed in the previous section should be adequate to assess whether it is substantial enough to affect a given indicator. The question here is not whether

accidental manipulation occurs but whether it is common and systematic enough to substantially alter the meaning of a given indicator.

A more problematic issue is the deliberate manipulation that may occur if alternative indicators are used to assess researchers when the researchers know about the choice of indicators in advance and have an interest in a positive outcome (Wouters and Costas, 2012). It is difficult to empirically evaluate the extent to which this is a problem for an indicator but it seems reasonable to assume that a web indictor will be deliberately manipulated whenever it can be. Although web indicators all fail this test, they can still be used for other types of evaluations. Also, if the evaluations are not of a high value nature, then it may be possible to employ strategies to reduce the likelihood of manipulation, such as honesty clauses or a degree of random or automatic checking of the data for signs of manipulation (e.g., Zimmermann, 2013).

2.3 SUMMARY AND RECOMMENDATIONS

The methods described in this chapter (correlation tests, creator interviews, or questionnaires, source content analysis and pragmatic evaluations) can all give evidence about the value or meaning of new indicators. While all of the methods have limitations, these can be at least partially overcome by using multiple different types (method triangulation). The following strategy is recommended for researchers seeking to evaluate any alternative indicator, based upon the above discussion.

1. Coverage analysis to assess the proportion of documents that have a non-zero score for the indicator. Low coverage restricts the number of practical applications of the indicator. In practice, coverage analyses are often conducted in parallel with correlation tests.

2. Correlations between the indicator and citation counts for diverse sets of individual fields and years to identify when they are likely to work. This is the simplest test to apply on a large scale and so is the first step. This also addresses the issue of whether the indicator works at all, but does not give evidence about the type of impact reflected by the indicator, if different from research impact. For this, at least one of the approaches below is needed unless the impact type is obvious.

3. Content analyses of selected sources of the indicator to find out why they were created. These are a logical next step because of the likely greater coverage in comparison to interviews and greater simplicity in comparison to surveys. The results will help to validate a claim for the type of impact that they may reflect (e.g., societal and educational). In addition, the results can help to give a finer-grained interpretation of the meaning of the indicator. Content analyses are not possible for usage indicators, which do not provide qualitative context.