The counting house

Scientists' work is often evaluated using citation statistics compiled by a company called the ISI. But how useful and reliable are the data? David Adam gets the measure of citation analysis.

There are, it is said, three types of lies: lies, damned lies and statistics. Many scientists, who find their work assessed through attempts to gauge how often it is cited in the scientific literature, would surely subscribe to that view. Citation analysis, in the hands of non-experts, can be an extremely blunt instrument. What's more, the specialists in the field have found that raw citation data often contain errors.

Practitioners of citation analysis rely on data gathered by the ISI, a company in Philadelphia formerly known as the Institute for Scientific Information. Over the past four decades, the ISI has scanned reference lists in scholarly publications and collated citations to previously published work. The resulting database was actually developed for information retrieval — allowing researchers to conduct rapid literature searches and to identify individual scientists working on particular topics.

Because it is hard for governments, funding agencies and promotions committees to find reliable yardsticks for measuring research quality, however, they often use the ISI's citation data to help them perform such evaluations. Important papers, the argument goes, will be cited more frequently. As a general rule, that is a reasonable assumption. But apply it blindly, without regard to the quality and limitations of the raw data, and the conclusions you draw may be far from reasonable.

"You must be very careful because it's about the reputation of scientists," says Anthony van Raan, director of the Centre for Science and Technology Studies at Leiden University in the Netherlands. "A database

prepared for information retrieval is not 100% suited to evaluation."

The ISI has always acknowledged the limits of citation analysis. But the lure of numbers has proved irresistible to those charged with judging scientists' work. And since the ISI was sold by its founder, Eugene Garfield, in the early 1990s, it has reacted to this demand. Now owned by the Thomson Corporation of Toronto, the company is producing software packages to help users probe its database, including one launched last year, called Essential Science Indicators, that promises to evaluate "potential employees, collaborators, reviewers, and peers".

At the same time, the ISI has toughened its attitude on the use of its data by independent bibliometrics researchers. As a result of Thomson's more hard-headed business

stance, some groups complain that they now face price rises and restrictions on data use that may force them out of citation analysis altogether.

Many bibliometrics researchers are concerned about this trend because they claim that the independent groups help to provide quality control for citation data. Others warn that the ISI is introducing products that are ripe for misuse. "They cover themselves by addressing all the issues in the accompanying notes, but who ever reads those?" asks one bibliometrics researcher. "It is frustrating for units such as ours."

Hit and miss

As an example of what can happen when citation statistics are misused, many experts point to the ISI's journal impact factors — a measure of the average number of citations garnered by the papers that each journal contains. Publishers have eagerly latched onto these data, using favourable impact factors in promotional material for their journals, and librarians have found them to be a convenient guide in deciding which journals to subscribe to.

But the use of journal impact factors has gone much further, extending to the evaluation of individual institutes, departments and scientists. The most obvious measure of the interest in various researchers' work would be to count citations to their papers directly. That is possible from the ISI's data, but getting the ISI or an independent group to conduct the analysis has historically proved expensive and time-consuming. So as a cheap-and-cheerful alternative, many evaluating bodies look at scientists' publication records and evaluate the quality of their output in terms of the impact factors of the journals in which their papers appear — figures that are readily available. It's "the poor man's citation analysis", says van Raan.
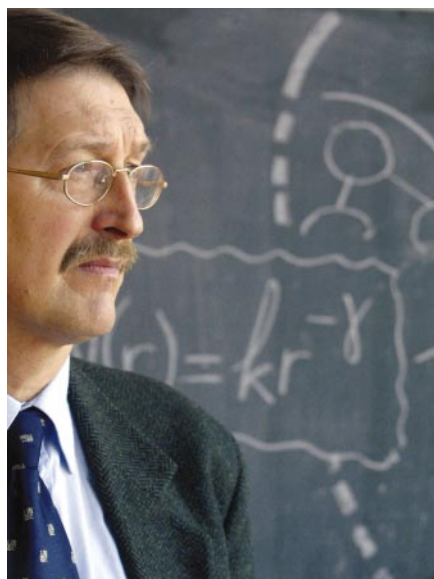
Anthony van Raan sees evaluations using journal impact factors as 'the poor man's citation analysis'.

Universities in Germany, for instance, regularly plug the impact factors of journals in which scientists publish into formulae to help them determine departmental funding. The Italian Association for Cancer Research requires grant applicants to complete worksheets calculating the average impact factor of the journals in which their publications appear. Elsewhere, the implicit use of journal impact factors by committees determining promotions and appointments is endemic.

Field trips

So why the controversy? Draw up a list of journals in a particular field and, with a few exceptions, there seems a pretty good correlation between a journal's impact factor and its perceived quality. But start making comparisons between fields — something the ISI warns against — and the results quickly become meaningless: mathematics researchers rarely cite more than one or two references, for example, whereas a typical paper in molecular biology includes dozens. This causes a wide variation in impact factors, even between comparable journals serving different disciplines, notes Per Seglen, a cancer researcher who also works on bibliometrics at the Norwegian Radium Hospital in Oslo, Norway (see chart, left). The figures are also biased in favour of journals that predominately publish review articles, which tend to be cited more frequently.

Other problems are less obvious, but when it comes to evaluating individual scientists, potentially more serious[1,2]. Seglen points out that about 15% of the articles in a typical journal account for half of the citations gained by that publication (see chart, right). This means that a typical paper in a journal with a high impact factor may

not, in fact, be cited much more frequently than the average paper in a lower-ranking journal. "There is a general correlation between article citation counts and journal impact, but this is a one-way relationship," Seglen says. "The journal does not help the article; it is the other way round."

Again, the ISI cautions against using journal impact factors to evaluate individuals. "These scores were never designed by ISI to be proxies for the influence of papers or, when aggregated, the work of individuals," says David Carter, the company's vice-president for corporate communications.

But that has not halted the trend, of which Finland provides an extreme example. There, government funding for university hospitals is partly based on publication points, with a sliding scale corresponding to the impact factor of the journals in which researchers publish their work.
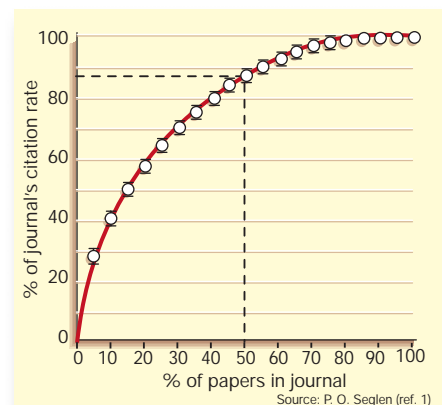
"To my knowledge, Finland is the only country in which the journal impact factor has been canonized in the law of the land," says Kari Raivio, rector of the University of Helsinki. Raivio calculates that a single paper published in a journal with an impact factor of 3, rather than 2, could have boosted a hospital's funding by about US$7,000 in 2000.

Uncertain ratio

The ISI calculates a journal's impact factor for a given year by searching its database for the number of citations that year to articles published in the journal in the preceding two years, and dividing by the number of 'citable' papers published by the journal in those two years. But this can raise another problem because the numerator in the equation can include citations to articles that do not appear in the list of citable items — generally restricted to original research papers and review articles. According to Henk Moed, a researcher at van Raan's Leiden centre, this can produce an impact factor that is up to 40% too high in some
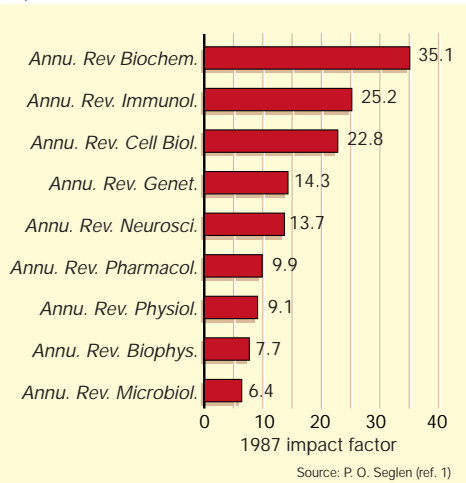
Different strokes: the impact factors vary between comparable journals in various disciplines.

Uneven split: about half the citable papers in a journal account for over 85% of its citations.
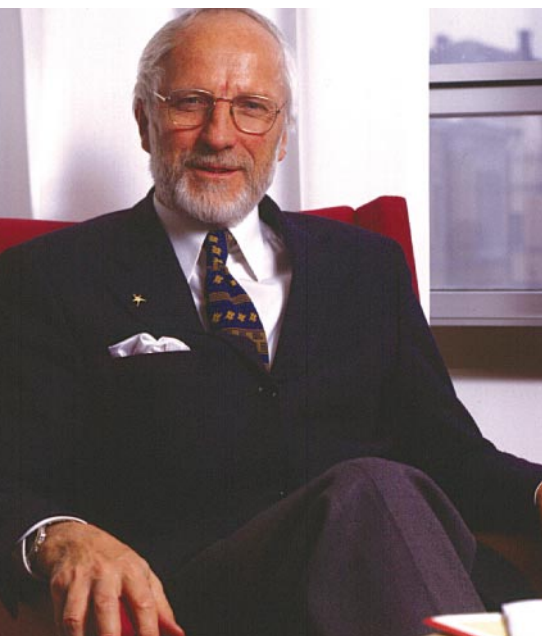
> F inland is the only country in which the journal impact factor has been canonized in law.
>
> **Kari Raivio**

title, volume, year and page number. Next, a computer takes a few bytes of information from each highlighted field to build up an identifying code or 'tag' that is unique to that paper. A similar data-capture and tagging process occurs for the references at the end of the paper. Algorithms then compare the citation tags with any article tags already in the database, and each successful match counts as a citation.

Journal impact factors are calculated using a simpler method. Of the various fields highlighted in each reference, only the title of the journal and the year are used. This generates a bulk count of references to a particular journal in any given year, but means that citations cannot be matched directly to individual articles. This helps to boost the impact factors of journals — such as *Nature*, *Science* and several leading medical journals — that include many different sections, including

news and correspondence pages, which may gain citations but are not 'citable' papers.

These biases are systematic, but investigations by *Nature* suggest that the ISI's journal impact factors can, on occasion, be subject to less predictable problems, arising from fluctuations in the counts of 'citable' items. For *Nature Genetics*, there appears to have been a significant undercount of citable items in 1996. More recently, the count of citable items for *Nature* in 2000 seems to have been inflated by the erroneous inclusion of items other than original research reports and review articles — including some from a section called Futures, which featured short science-fiction stories.

Blazick admits that some mistakes are possible as the 'citable' and 'non-citable' articles must be separated manually. "There's some room for error with how fast we do the process," she says, claiming that errors are likely to even out, with as many citable articles missed as non-citable ones included.

The ISI declined to comment on specific cases, but its officials say that the company does address errors that are drawn to its attention and is always striving to improve accuracy. "We are constantly coordinating with our publishers to enhance our processes," says Carter.

One currently live issue is the development of procedures to collate citations to

cases. To understand why, it is necessary to know how the ISI counts citations.

Pamela Blazick, editor of the ISI's *Journal Citation Reports*, in which journal impact factors appear, says that pages from journals are first scanned using optical character-recognition software. To store a research paper in its database, ISI employees highlight the following fields: author, address, journal

Pretenders to the throne

The ISI is the undisputed king of the counting houses, but new software that scans articles, pulls out references and automatically generates citation indexes may challenge its monopoly, as publishers and digital libraries start to adopt these tools.

Like the ISI's database, these new products have arisen from efforts to improve the accessibility of the academic literature. CrossRef, for instance, is a non-profit consortium — to which 99 publishers have signed up — that allows users to click straight from the references in online papers to the abstracts or full texts of some 3.9 million research articles in more than 5,600 journals.

To achieve such a set-up requires software, manual systems or a mixture of the two that can accurately identify all references — which will be written in a variety of formats and are often replete with mistakes. Such systems must also identify elements such as author names and starting page numbers. Each reference is then compared with those in a central

database to find an identifier — a URL for the web page of the paper concerned, or the paper's unique digital object identifier (DOI).

But once references and papers are interlinked, it is relatively simple to apply conventional algorithms to create citation indexes, impact factors and other indices, and to rank cited researchers and articles, much as the ISI has been doing for years.

CrossRef has so far made no moves in this direction — but other groups have. Among the pioneers are the developers of ResearchIndex at the NEC Research Institute in Princeton, New Jersey. This tool has been used to build, without human intervention, a digital library of more than 300,000 papers on computer science gathered from the web. ResearchIndex is not error-free — in tests, it has matched references to articles with about 95% accuracy; for journals with standard citation practices, its performance should be close to 100%. But project leader Steve Lawrence argues that the system could be expanded to cover a wider range of publications than the

ISI's database, which is more cost-intensive because of its manual inputs.

Fully automated citation-indexing systems are proliferating. The Open Citation Project is building one for electronic preprint archives, and NASA's Astrophysics Data System (ADS) has one for its collection of 3 million abstracts. Günther Eichhorn of the Smithsonian Astrophysical Observatory in Cambridge, Massachusetts, project scientist for the ADS, says that users are increasingly using the system to create their own citation statistics.

To improve results, companies such as Parity Computing of San Diego are using a hybrid approach, automating most functions, but also employing people at key points to increase quality control. Chris Rosin, Parity's president, claims that its system achieves over 99% accuracy. The company recently implemented a citation index for the 50-year literature archive of the Association for Computing Machinery, the world's largest computing society. Parity intends to build on such citation

indexes to generate impact factors and other citation statistics, says Rosin.

None of these initiatives has yet attempted the multidisciplinary coverage offered by the ISI's database. And given the ISI's dominant market position, it remains unclear whether anyone will. "Even with automation there are high costs of obtaining and organizing a massive amount of journal literature in usable form, and doing quality control," Rosin says. **Declan Butler**

CrossRef
➧ www.crossref.org
ResearchIndex
➧ citeseer.nj.nec.com/cs
Open Citation Project
➧ opcit.eprints.org
Astrophysics Data System
➧ adswww.harvard.edu
Parity Computing
➧ www.cparity.com

Declan Butler is *Nature*'s European correspondent, and a member of a CrossRef working group exploring search technologies.

papers authored by consortia, rather than a conventional list of individuals. Citations to such papers appear to have been under-counted by the ISI. This came to public attention in January[3,4], after *Nature* investigated the suspiciously low citation count for last year's landmark paper on the draft human genome sequence[5], from the International Human Genome Sequencing Consortium. As it turned out, the ISI was only considering citations to the full list of authors, led by Eric Lander of the Whitehead Institute for Biomedical Research in Cambridge, Massachusetts.

Warped records

Nature is not the first to point out irregularities in the ISI's data[6,7] (see Correspondence, pages 731–732). Individual researchers have also found errors after probing records for their own publications held by the ISI. Perhaps the largest source of error is the tendency of scientists to make mistakes when citing one another's work. Rather like the attribution of the famous 'lies' quote used to open this article — often associated with Mark Twain but in fact borrowed by the writer from the Victorian-era British Prime Minister Benjamin Disraeli — errors in references mean that citation statistics can misplace credit. Indeed, errors can creep in to every data field recorded by the ISI's database. And as the citation counts creep up, so do the errors as scientists copy them from paper to paper.

Indeed, experts in bibliometric analysis say that, for particularly highly cited papers, it is not unusual to find that variants with, say, an error in the journal volume number have themselves garnered enough citations to beat the vast majority of papers in the ISI's database. Variations in addresses also cause havoc, especially with national, regional and institutional comparisons.

Given these difficulties, and the huge volume of data processed by the ISI — covering some 5,700 science journals — bibliometric experts concede that the company faces an extremely difficult task. The ISI has a quality-control department to clean up the data, and says that the algorithms that match articles have been refined to cope with common typing errors, including misspelt names. But the output quality remains at the mercy of variations in the input. "Nobody knows how accurate the raw data are, but they're certainly not clean, accurate data in the way that some people think," says Ben Martin, a bibliometrics expert at SPRU, the unit for science and technology policy research at the University of Sussex in Brighton, UK.

When specialists such as Martin analyse the ISI's data, they go to great lengths to remove erroneous citations. But anyone who purchases the Essential Science Indicators software can now probe the information in the company's databases to evaluate institutions, departments and individuals.

Although products such as Essential Science Indicators will allow non-specialists to avoid the mistake of using journal impact factors as a proxy for direct citation counts, many experts are concerned that they will enhance the weight given to citation analysis by evaluating committees. And given that non-experts are generally not aware of the problems caused by errors in the data, this could have serious consequences.

"The data these products produce at the level of the institution or individual can be very 'noisy' and require a considerable amount of cleaning before they can legitimately be used by policy-makers and analysts," says Linda Butler, a bibliometrics researcher at the Australian National University in Canberra. She cites an extreme example: the Walter and Eliza Hall Institute of Medical Research (WEHI), a leading medical facility located in the grounds of the Royal Melbourne Hospital. The hospital appears in the first line of the WEHI address and so collects up to 70% of the citations intended for the independent institute in the subject listings in Essential Science Indicators.

The ISI makes such problems clear in the product notes that accompany the software. "But I know from considerable experience that researchers, analysts and bureaucrats will head straight to the data and start playing with them without any real understanding of their quirks or deficiencies," Butler says.

Linda Butler: data can need a lot of cleaning.

At the same time as launching 'do-it-yourself' citation-analysis software, the ISI has taken a tougher stance on independent research groups that want to use its data. One bibliometrics researcher, who did not wish to be identified, claims that the cost of buying certain ISI data has risen almost fourfold since 1995.

Tibor Braun, a chemist at Loránd Eötvös University in Budapest, Hungary, and editor of the journal *Scientometrics*, calls the era before Thomson bought the company the ISI's "romantic period". Garfield's personal interest in bibliometric research, observes Braun, meant that specialists in the field were given freedom to play with the data. "He pursued many things that were perhaps not essentially business concerns or cost-effective because he owned the company and he was interested in the results," agrees David Pendlebury, an ISI analyst.

The ISI's senior management declined to comment on the company's current business

Nobody knows how accurate the raw data are, but they're certainly not clean, accurate data in the way that some people think.

strategy. But few would dispute that the ISI is acting within its rights in exerting tighter control over its database, which is a valuable asset. For the time being, the ISI retains an effective monopoly in the field of multi-disciplinary citation data — although advances in information technology may yield alternative databases (see 'Pretenders to the throne', opposite).

Mining the business

Braun says that some researchers exploited the freedom they enjoyed in the Garfield era, developing lucrative businesses on the back of the ISI's data. "They saw a lot of people using their database, making products from the information and selling them," says Braun. The ISI's new management may have replaced romanticism with hard-nosed capitalism, says Braun, "but we should not condemn the capitalists".

Garfield confirms that piracy of the ISI's data was always a problem for the company. "From the outset, there were people who felt they could use the data without limit and without cost," he says. "Not the least of these were government bureaucrats who thought nothing of awarding contracts to companies other than ISI to perform studies using ISI data."

Although the ISI's business strategy may be perfectly legitimate, that leaves the bigger question of whether it, and citation analysis more generally, is good for science. Supporters of citation analysis argue that it injects objectivity into decision-making that can otherwise be rife with cronyism. Detractors counter that the practice is so riddled with errors and biases that it can be worse than useless.

As long as citation analysis continues to be used for scientific evaluation, this debate seems sure to continue — and you can cite us on that. ■

David Adam is a news and features writer for Nature.

1. Seglen, P. O. *Br. Med. J.* **314**, 498–502 (1997).
2. *Nature Neurosci.* **1**, 641–642 (1998).
3. Cherfas, J. *Science Watch* **13**(1), 8 (2002).
4. *Nature* **415**, 101 (2002).
5. International Human Genome Sequencing Consortium *Nature* **409**, 860–921 (2001).
6. Moed, H. F. & van Leeuwen, Th. N. *J. Am. Soc. Inf. Sci.* **46**, 461–467 (1995).
7. Reedijk, J. *New J. Chem.* **22**, 767–770 (1998).

➧ **www.isinet.com**