



THE UK'S SHARE OF WORLD RESEARCH OUTPUT

An investigation of different data sources and time trends

A report by the Research Information Network

June 2009



www.rin.ac.uk

The Research Information Network acknowledges the work of Grant Lewison and colleagues in the CIBER group at University College London in conducting this short study and in co-authoring the report.

This booklet is available online at www.rin.ac.uk/uk_presence_research

© Research Information Network 2009

Contents

1. Introduction	5
2. Sources of data	6
3. Counting methods	8
4. Year to be counted	9
5. Reports, sources and methodologies summarised	10
6. The impact of using different data sources and methodologies	11
7. Conclusion	19
8. Recommendations	20

Foreword

Bibliometrics have come to play an increasing role in assessing the performance of researchers in the UK, as indeed in other parts of the world. But the complexities of both the data sources and the methods of analysis used in bibliometrics are little understood by many of those who wish to make use of the results.

Some key features of these complexities can be seen in the published reports of the UK's percentage presence in world science. In the course of the Research Information Network's (RIN) work over the past few years, we have noticed that the figures provided in various reports for the UK's share of the world's production of scientific publications vary enormously. That a seemingly straightforward figure should show such volatility perplexed us, and so we asked Grant Lewison and colleagues in the CIBER group at University College London to analyse the published figures, and explain the differences between them. This report is the result of their work.

We believe the report highlights important issues both for those who produce and supply bibliometric analyses of research performance and for those who commission and make use of such work. Producers and suppliers must make much more transparent the choices they have made as to data sources and methodology, and the implications of those choices. Policy-makers and others interested in the health of the UK research base must also take greater care to interrogate and understand the figures that they use. Otherwise the risk is that policy and related decisions will be made on the basis of false assessments of performance.

We hope therefore that our report will help all those who have an interest in the performance of researchers in the UK to gain clearer understanding of the questions they must ask when producing or making use of even seemingly simple bibliometric analyses.

1. Introduction



Governments all over the world have become more interested over the past decade in how they can evaluate the performance of the scientists they support. Increasingly, they seek to do so by using quantitative measures of research publications, or bibliometrics. The prime measures are the numbers of papers published each year and the numbers of citations to those papers. Time series of these values are often used in the statistical reports on national scientific performance, and on most measures the UK is shown as second only to the United States in the volume both of its publications and of citations.

Like all statistical claims, the values and trends produced in the various reports depend on the sources of data and on the methodology used to generate them. This paper focuses solely on measures of the global volume of publications, and the UK's percentage share of that volume. It does not seek to analyse the more complex issues surrounding the measurement of citations. Nevertheless, it is striking that even in what might seem the more straightforward task of measuring publications as distinct from citations, there are huge variances in assessments of the UK's share. The figures given in different published reports for what we term the UK's 'percentage presence in world science' vary by 40%: figures of 6.5% and 9.1% have been shown for the single year 2002, for example, and there are significant differences in the shape of the trend line since then. With such large differences, it is difficult for policy-makers and others concerned with the health of the UK research base to get a clear picture of how well it is performing. This paper explains how the differences arise, and reflects on the implications for the measurement of UK scientific performance.

2. Sources of data

The key sources of bibliometric data underlying published reports up to now have been three databases of scholarly journals established by the former Institute for Scientific Information (ISI), now the Scientific business of Thomson Reuters: the Science Citation Index (SCI), the Social Sciences Citation Index (SSCI), and the Arts and Humanities Citation Index (AHCI). There is some overlap between the three indexes, and SCI is much the largest. The strength of the ISI databases is that for many years they were the only ones available that included the references in the published papers, and author addresses. It should be noted, however, that the exclusion from the SSCI and the AHCI of outputs such as monographs and chapters in edited collections, together with their focus on outputs written in the English language, means that they are not regarded as especially representative of the global output of researchers in the social sciences, arts and humanities. These limitations are not so important in the SCI, since in the physical and life sciences articles written in English are much more the norm.

In 2004, Elsevier launched a rival database called SCOPUS, with references going back to 1996. It includes outputs such as monographs, although the numbers are small. The basic database has been processed by the SCImago analysis group at the University of Granada, which has created a free web-based interface that enables SCOPUS to be interrogated to yield values for percentage presence in world science. As yet, however, SCOPUS has not been used as the basis for any published reports.

The figures given in all the major published reports on national scientific performance, therefore, have all been based on the SCI, SSCI and AHCI. These reports include:

- a report by the **Wellcome Trust** in 1998 on the percentage share of science publications for G7 countries, 1987-95
- annual reports produced by **Evidence Ltd** for the UK government. The latest, dated July 2008, includes a report on UK output in 2007
- biennial Science and Engineering Indicators reports published by the **US National Science Foundation** (NSF). The latest, dated 2008, gives bibliometric data for 2005 and earlier years

- annual or biennial reports on European science published by the **European Commission** (EC), including bibliometric data prepared by the Centre for Science and Technology Studies (CWTS) at the University of Leiden. The latest, dated 2007, gives bibliometric data for 2004
- biennial reports published by the French **Observatoire des Sciences et des Technologies** (OST). The latest, dated 2006, gives bibliometric data for 2002-04, and
- a report published by the **Research Council of Norway** (Forskingsradet) in 2007, with bibliometric data for 2006.

There are significant differences, however, in the precise data from Thomson Reuters on which these reports are based. The differences are of three kinds.

First, the data can be taken either from the 'full version' available on the Web of Science (WoS) or from CD-ROMs, which each cover a single year. The three indexes – SCI, SSCI, and AHCI – are each available in both versions. The WoS version has wider coverage than the CD-ROMs, mostly accounted for by its including journals published in languages other than English, and from countries outside the main publishing nations (the USA, UK, Germany and the Netherlands).

Secondly, some reports use data only from the SCI, some from the SCI and SSCI, and some from all three indexes.

Thirdly, some reports use data only relating to articles (including from 1996 'notes') and reviews, while others include letters and conference proceedings (which are important in some subject areas such as engineering, but not in others).

3. Counting methods



Two principal methodologies are used to count percentage presence, termed integer counting and fractional counting. The differences between the two have become increasingly significant as international collaborations, resulting in multi-authored papers, have become more common.

In **integer** counting, a paper with two authors from a UK address and one from a French address would be counted as one for each country. In **fractional** counting, the paper would count as 0.67 for the UK and 0.33 for France. Some variants of fractional counting are occasionally used, for example by giving additional credit to the first (or last) author, but these are not used in national statistics because authorship practices vary in different fields.

It can readily be seen that a country's presence based on integer counting is bound to be greater than its presence based on fractional counting. So countries with high levels of international collaboration (typically ones with a smaller scientific output) will show the biggest difference in the two measures of percentage presence. It is important also to stress that when based on integer counting, statements of the kind 'the UK produces x per cent of the world's scientific publications' are at best misleading. For the 'percentage' figure given is a numerator which should be linked to a denominator greater than 100.

There is a third **hybrid** counting method, which uses integer counts for individual countries' outputs, but calculates as the world total (the denominator) the sum of the individual country totals.

4. Year to be counted

Web of Science uses two different definitions of the year in which publications are to be counted:

- publication year: the year when an article appeared in print (which may be different from the year it first appeared online), and
- database or campaign year: the year when the publication was added to the index. A given database year normally contains about 10% of the papers from the previous publication year, and a small percentage also from earlier years and from the subsequent publication year (papers available online before the print version is available).

5. Reports, sources, and methodologies summarised

Table 1 shows the data sources and the methodologies used in each of the six reports listed above, along with those used by SCImago, based on SCOPUS data.

Table 1. Sources and methods used in seven reports giving the UK percentage presence in world science

Report	Data source*	Papers included†	Year	Counting method
Evidence Ltd	SCI, SSCI, AHCI	A, R	Database	Integer
Forskningsradet	SCI, SSCI	A, R	Database	Hybrid
NSF	SCI, SSCI (CD)	A, R	Publication	Fractional
OST	SCI	A, R, L	Publication	Fractional
SCImago	SCOPUS	A, R	Publication	Integer
EC	SCI, SSCI	A, R, L	Database	Integer
Wellcome Trust	SCI (CD)	A, R	Database	Integer

* Values are taken from the Web of Science unless marked (CD), when they are taken from the CD-ROM. SCOPUS includes documents in science, social science, and arts and humanities.

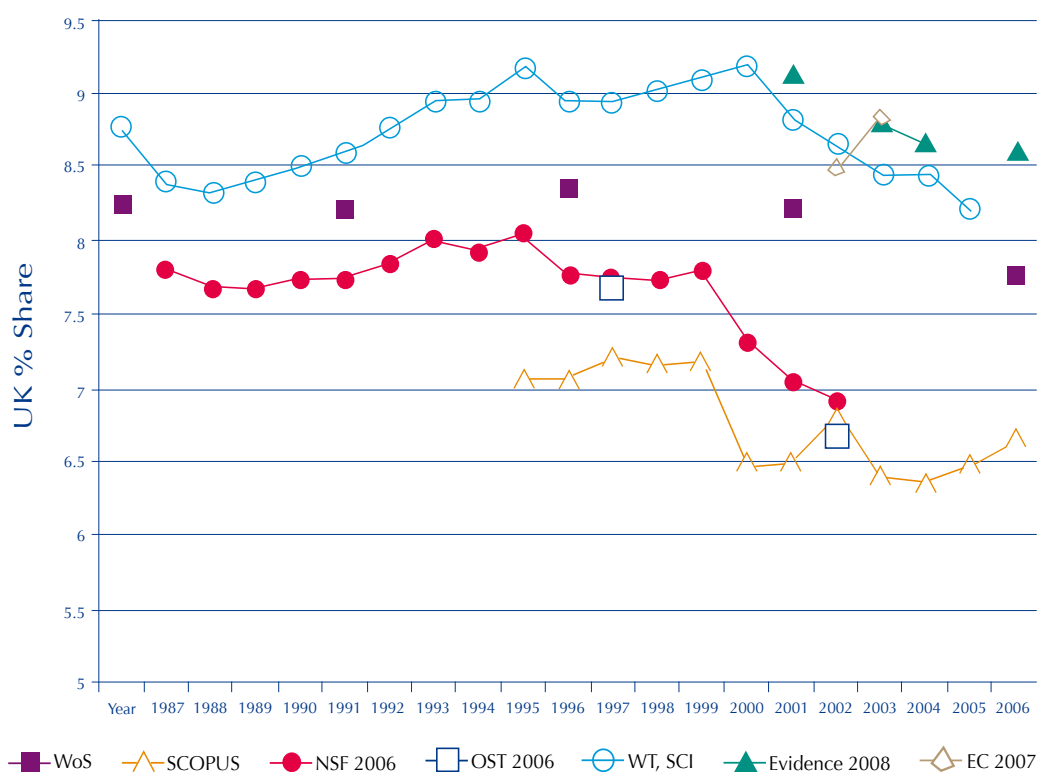
† A = articles (includes notes > 1996), R = reviews, L = Letters.

6. The impact of using different data sources and methodologies

6.1 Differences between the published figures

As a result of the differences in the data sources and the methodologies used in the published reports, the values shown for the UK percentage presence vary significantly, as is shown in Figure 1.

Figure 1. Values shown for UK percentage presence in different reports



There is a clustering of values around two different trend lines. But even in 2007, the latest year for which there are publicly-available estimates, and when there appears to have been some convergence, the UK percentage presence varies between 6.7% and 8.3%. With such large variations in the values shown in different reports, it is difficult to get a clear picture of UK research performance. In the rest of this paper, we seek to explain in more detail how these differences arise, and how the figures might be corrected to make them more consistent.

6.2 Data sources and types of publication

The SCI is the predominant index: for the publication year 2007 the SCI contained 91% of all the articles in the combined indexes, the SSCI 10% and the AHCI 3% (the total is greater than 100% because there is some overlap between the indexes). Table 2 shows the UK percentage presence for the publication year 2007, using integer counts for each of the three databases.

Table 2. UK percentage presence (integer count) in publication year 2007 in the SCI, SSCI and AHCI for different publication types.

Source	Articles (A)	Reviews (R)	Letters (L)	A + R	A + R + L
SCI	7.51	12.45	12.54	7.77	7.95
SSCI	13.10	17.52	7.54	13.36	13.17
AHCI	9.07	16.65	8.40	9.46	9.39
SCI+SSCI	7.85	12.66	12.21	8.10	8.25
SCI+SSCI+AHCI	7.85	12.68	12.02	8.09	8.24

Clearly, there are big differences between the UK percentage presence shown in the three indexes, and for different categories of publication. Thus the SSCI shows the highest UK percentage presence both for articles and overall, but the lowest presence for letters. The numerical dominance of the SCI, however, means that the strong UK presence in the other two indexes is much diluted when all three indexes are aggregated, whether counts are of articles and reviews only, or of all three document categories.

The key figure in Table 2 is probably the 7.77% shown as the UK presence in articles and reviews in the SCI. The table shows that the inclusion of letters, which seldom present significant research findings, inflates the UK presence in the SCI by about 0.18%, and in the aggregate of all three indexes by about 0.15%. The Anglophone bias of the SSCI and the AHCI probably inflate the UK presence in the aggregate of all three indexes by a further 0.32%.

A further set of differences arises from the use of either the Web of Science or the CD-ROM (since 2004 DVD) versions of the SCI. Table 3 shows the UK percentage presence for articles, notes and reviews in the two versions of the SCI (again using an integer count) in four sample years.

Table 3. UK percentage presence (integer count) of articles, notes and reviews in the SCI in four database years

Year	WoS version	CD-ROM version	Difference	UK presence in journals not in the CD-ROM
1987	8.26	8.77	0.51	5.39
1992	8.24	8.61	0.37	6.10
1997	8.36	8.95	0.59	6.07
2002	8.23	8.66	0.44	6.65

It is clear from the table that use of the CD-ROM version inflates the UK presence by an amount varying from 0.37% to 0.59%. This arises because, as noted above, the WoS version of the database includes more non-English-language material, and more published outside the US, UK, Germany and the Netherlands. The difference between the two figures varies from year to year, and there is no discernable time trend; but on average it is 0.48%.

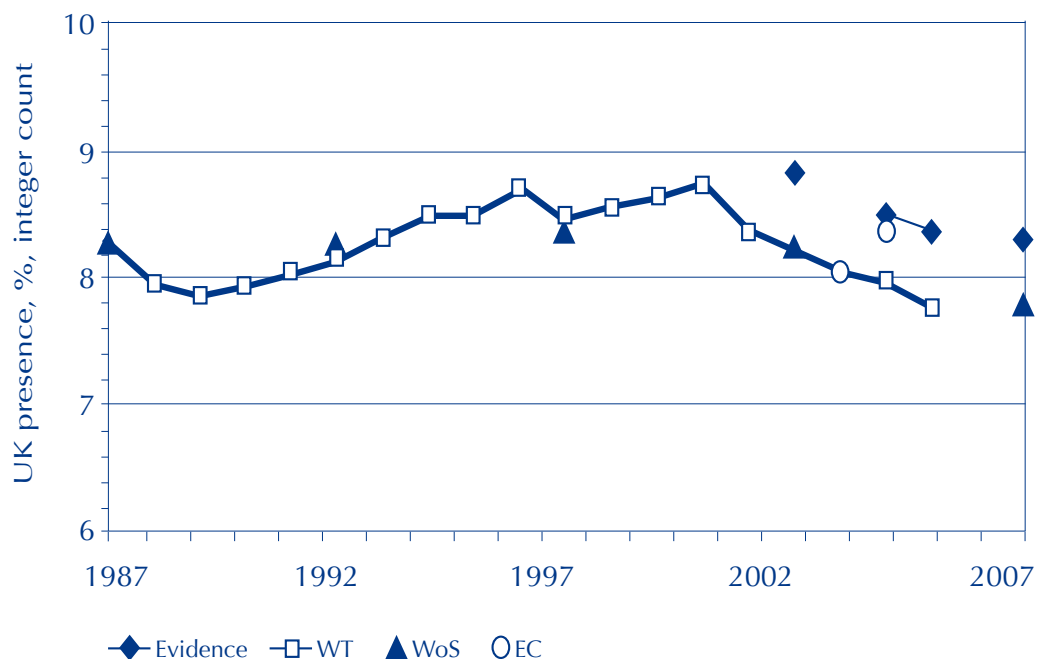
In summary there are thus three possible corrections to apply, if the UK percentage presence is not to be overstated:

- a correction of -0.18% to counts that include letters, on the grounds that these seldom contain substantive research findings;
- a correction of -0.32% to composite counts including the SSCI and the AHCI, to reflect the Anglophone bias of those indexes; and
- a correction of -0.48% to counts based on the CD-ROMs, to reflect the relative absence from those counts of non-English-language and other material.

6.3 Integer and fractional counts

Figure 2 shows values for UK percentage presence from those reports listed in Section 2 above that are based on integer counts. The Wellcome Trust values were obtained for years after 1995 using the methodology employed in the 1998 report. All values are corrected to take account as relevant of the factors set out in Section 6.2 above.

Figure 2. Corrected estimates of the UK percentage presence in the SCI, 1987-2007, using integer counts

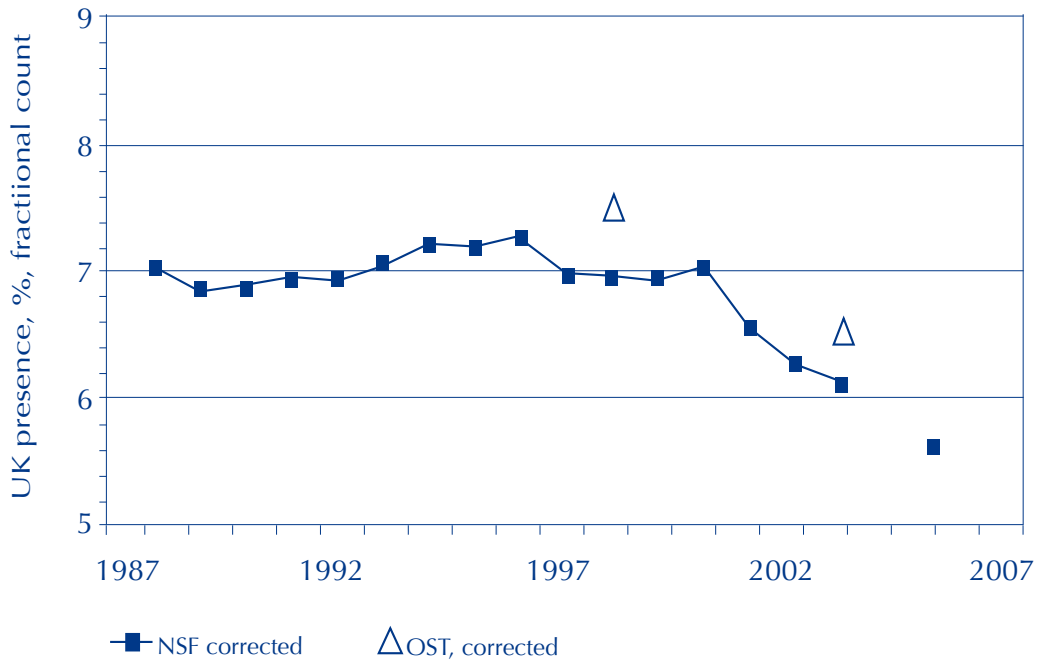


WT = Wellcome Trust, corrected to WoS values. Evidence = Evidence UK, corrected to SCI only values. EC = EC/Leiden, corrected to exclude letters and to SCI only values.

It is important to stress that the values shown here for UK percentage presence are lower than those in the published reports, because they have been corrected to take account as appropriate of factors such as Anglophone bias in the SSCI and AHCI. The values shown for Evidence, therefore, have been reduced by 0.32% from those in its published reports. When corrected in this way, the figures, and the patterns shown are broadly consistent at around 8%, although the Evidence figures, and the EC/Leiden figure for 2004, are still slightly higher than those shown by the Wellcome Trust and Web of Science. All except EC/Leiden show a decline in the UK percentage presence in the past five-six years. This decline – to be seen also in calculations of the US presence – is attributable mainly to the rapid rise in the numbers of papers from China and other Asian countries.

Figure 3 shows the values for UK percentage presence from those reports listed in Section 2 that are based on fractional counts. Again, the figures are corrected to take account of the factors listed in Section 6.2.

Figure 3. Corrected estimates of UK percentage presence in the SCI, 1987-2007, using fractional counts



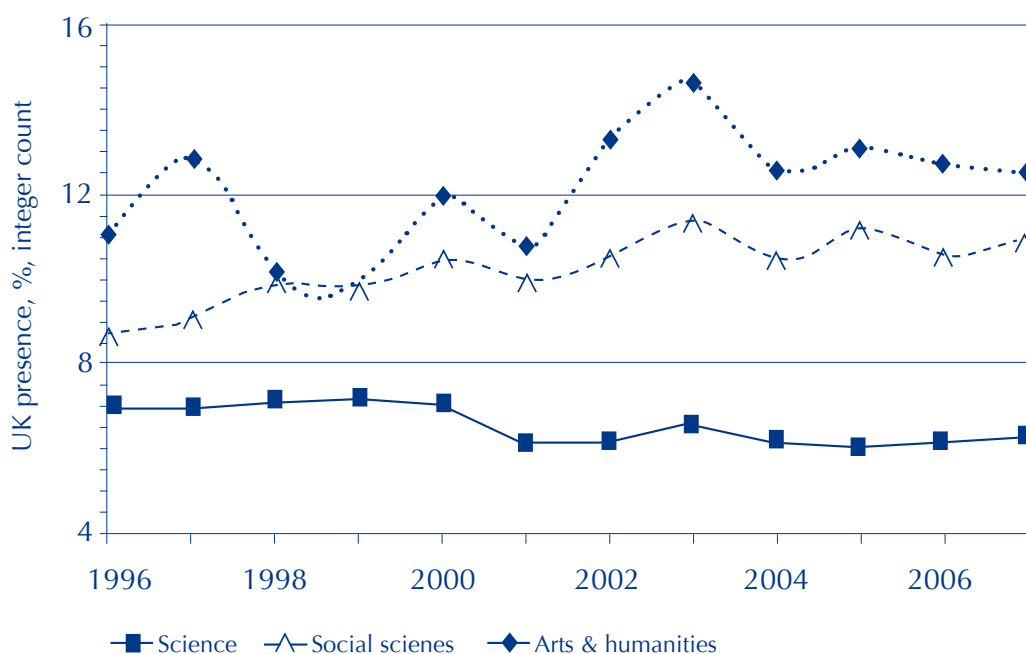
NSF = US National Science Foundation (data from 2006 and 2008 reports).
 OST = French Observatoire des Sciences et Technologies. Values corrected to WoS SCI only values.

Again, the figures for UK percentage presence shown here are lower than in the published reports, to take account of factors such as the use of the ISI CD-ROM and the inclusion of letters. When corrected in this way, the values and patterns shown are similar, this time at around 7%. The decline from 2000 is sharper, however, as a result of the rise of international collaboration.

6.4 The SCOPUS database

The values calculated from the SCOPUS database as presented by SCImago are based on integer counts and are shown in Figure 4. The figures cannot be corrected in the same way as those based on SCI, SSCI and AHCI, since they are based on different source data. They show a similar pattern of a higher UK percentage presence in social sciences, arts and humanities than in science. But it should be noted that the non-science fields represent a lower proportion of the total database than with Thomson Reuters: social sciences represent about 4.2% of the documents, and arts and humanities about 0.4%.

Figure 4. UK percentage presence (integer count) in the SCOPUS database, as shown by SCImago

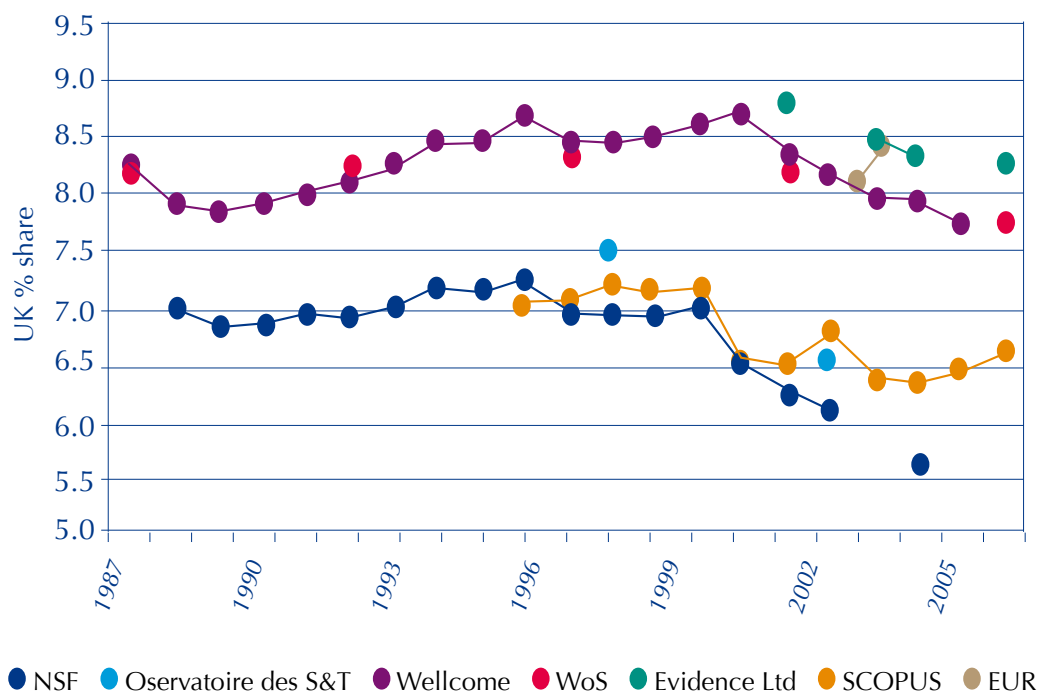


The values shown for science are lower than those using corrected integer counts based on the SCI. The pattern over time shows some similarities, with a decline from a peak in 2000; but the decline is much less sharp. This is partly because SCOPUS has a higher proportion of Chinese documents (over 12% in 2007, almost twice the UK figure) than WoS (10.2%). But the main difference probably arises simply from the range of journals selected by SCOPUS as compared with WoS.

6.5. Corrected integer and fractional counts compared

For ease of comparison, figure 5 shows the corrected figures for integer and fractional counts together.

Figure 5. Corrected Values for UK percentage presence shown in different reports



The chart shows clearly that integer counting gives higher values for the UK percentage presence than fractional counting. This is partly because, as noted above, the 'percentage' figure is in fact a numerator that should be linked to a denominator greater than 100. It is also clear, however, that even when corrections are made to take account of different data sources and methodology, there remain unexplained differences of around 0.5% between the values derived from the two different methods: for integer counts, the Evidence figures are about 0.5% higher than the WOS/Wellcome Trust figures; and for fractional counts the OST figures are about 0.5% higher than the NSF figures.

7. Conclusion

Bibliometrics are playing an increasingly important role in the assessment of research performance, at national and international as well as at institutional and individual levels. As we have shown, however, the decisions that are made in choosing the sources from which to count, and the methods of counting, make for very different results. It is critically important that the decisions made should be transparently recorded in published reports, and that they should be opened to reasoned debate. It is not acceptable that 40% differences in the published values for the UK percentage share of global scientific publications for a single year should be left unexplained.

We do not suggest that there is a single right approach in addressing the issues raised in this report; and there is no single definitive figure to represent the UK percentage presence in world science. It depends on what you want to count, and how, and (most important) why. A first requirement is to be clear as to the question to be answered when undertaking, or commissioning, a report using bibliometric data of the kind considered in this report. For there are important risks in using bibliometric measures to assess performance and to inform policy-making if insufficient effort is made to record and understand the implications of the methodological choices made. This becomes the more important as new sources of data and new tools for manipulating them become available.

Recommendations

We therefore recommend that:

1. Producers and publishers of bibliometric data should ensure that they explain as precisely as possible the nature and the implications of the choices they have made as to their sources and methods, and wherever possible how and why they differ from those used and published elsewhere.
2. Policy-makers and others who commission or make use of reports based on bibliometrics should interrogate their suppliers as to the data sources and methods that are employed, and as to the implications of the choices that are made.
3. All those who supply bibliometric analyses, and those who publish and make use of them, should take great care in how they present their findings, to ensure that they avoid misleading statements as to the nature and scope of their measurements. They should take particular care with general statements about proportions of the world's "scientific publications", and with the framing of any statements based on integer counts.

References

Dawson G., B. Lucocq, R. Cottrell and G. Lewison (1998) *Mapping the Landscape: UK biomedical research outputs, 1988-95* (The Wellcome Trust, London: Policy Report no 9)

Department for Innovation, Universities and Skills (2008) *International comparative performance of the UK research base* (Report by Evidence Ltd, July 2008)

European Commission. *Towards a European Research Area Science, Technology and Innovation: Key Figures 2007* (Luxembourg: Office for Publications of the European Communities)

National Science Board (2008) *Science and Engineering Indicators 2008*. Two volumes (Arlington, VA: National Science Foundation)

Observatoire des Sciences et des Techniques (2006) *Indicateurs de sciences et de technologies, Édition 2006* (Paris: Éditions Economica & OST)

Research Council of Norway (2008) *Report on Science & Technology Indicators for Norway 2007* (Oslo, February 2008)

Glossary

AHCI Arts and Humanities Citation Index

CIBER Centre for Information Behaviour and the Evaluation of Research

CWTS Centre for Science and Technology Studies

EC European Commission

ISI Institute for Scientific Information

NSF National Science Foundation

OST Observatoire des Sciences and Technologies

RIN Research Information Network

SCI Science Citation Index

SSCI Social Sciences Citation Index

WoS Web of Science

Research Information Network

The Research Information Network was set up in 2005 by the four UK higher education funding bodies, the seven research councils and the three national libraries. Our role is to enhance and broaden understanding of the information resources and services available to researchers, and how they use them; and to promote innovation and the development of effective policies and strategies for the benefit of the UK research community. www.rin.ac.uk

Photo credits: cover © istockphoto, p5 ©
Doug Vernimmen, p8 © istockphoto

Get in touch with us:

Research Information Network
96 Euston Road
London
NW1 2DB

T +44 (0)20 7412 7964

F +44 (0)20 7412 7339

E contact@rin.ac.uk

www.rin.ac.uk