


Subject: Library and Information Science

Production of Courseware

 **-Content for Post Graduate Courses**



Paper No : 10 Informetrics and Scientometrics

Module : 11 Science Indicators



Development Team

Principal Investigator
&
Subject Coordinator

Dr. Jagdish Arora, Director
INFLIBNET Centre, Gandhinagar

Paper Coordinator

Dr I K Ravichandra Rao
Retd Professor, Documentation Research and
Training Centre

Content Writer

Dr Sujit Bhattacharya
Senior Principal Scientist, CSIR- NISTADS

Content Reviewer

Dr I K Ravichandra Rao
Retd Professor, Documentation Research and
Training Centre

Unit 11

Science Indicators

I. Objectives

The students after reading this module and doing the exercise should obtain basic understanding of indicators in general and S&T indicators in particular. They would also be able to apply indicators for measuring some facets of scientific activity.

II. Learning Outcome

At the end of this module, you have gained knowledge related to various indicators and their merits; you have also learnt -- how to compute? How to interpret? What are the various limitations of various indicators? Etc. You have now learnt an important chapter in scientometrics; this will be very useful in carrying out research projects in the area of National mapping of Science.

III. Module Structure

1. Introduction
2. Science and Technology Indicators
3. Input Indicators
4. What is Scientometrics?
5. Scientometric Techniques
6. Science and Technology Performance Indicators
7. Identifying Conceptual Connections among Documents
8. Co-Citation Analysis
 - 8.1 Similarity through identifying jointly cited papers (Co-Citation)
9. Co-word Analysis
10. Methodological problems of bibliometric based indicator
11. Summary
12. References

1. Introduction

Indicators are used to measure the various Dimensions that are perceived to constitute a Construct. Thus there are two important concepts namely dimension and construct that requires understanding. Constructs can be thought of an entity that cannot be directly measurable. For example creativity, performance, and intelligence are constructs. They differ from variables such as weight, blood pressure, and temperature that can be measured on a scale. Construct is thus not a single measurable entity but can be expressed through measurement of directly observable variables. Constructs are composed of various dimensions. For example financial indicators can show the health of a country's economy. Science and Technology indicators help to capture various facets/characteristics of science and technology such as productivity, collaboration, impact.

An indicator should convey information about a particular element or a sub-element that it represents. Indicators are based on statistics covering various aspects of the phenomena. An ideal indicator should be representative — it should cover the most important aspects of the elements concerned. It should be reliable — in that it should directly reflect how far the objective concerned is met, well founded, accurate, measured in a standardized way; and feasible — data should be readily available, and at reasonable cost.

Construction of proper indicators is dependent on Reliability and Validity. Reliability implies consistency of measurement i.e. an indicator is reliable if different people who use them get consistent results. Validity is concerned with the accuracy of the measurement i.e. indicator should be able to measure what they are intended to measure. How appropriately proxies measure the various parameters provide validity to the indicators? For example indicators constructed from research papers are commonly used as proxy for measuring scientific activity. There is a strong rationale for choice of indicators based on research papers as proxy. However, research papers will not generally be published in areas of strategic/military research and where research has potentiality for commercial exploitation. In those cases, research papers would not be a proper proxy for measuring scientific activity. The database one is using to capture the research activity in a field should systematically cover all the important journals in that field. This ensures that data for the research field is reliable.

Important steps for measuring the construct through indicators: Subdivide the construct according to several set of dimensions. Create indicators for measuring each of the dimensions. Create the composite indicator that captures all the dimensions. This composite indicator is the construct. Each dimension in itself reveals important aspects of phenomena and thus in many cases we are interested to measure the dimension only.

2. Science and Technology Indicators

Science and Technology indicators help to capture various facets/characteristics of science and technology. Ideally they should describe the science and technology system, enabling better understanding of its structure, of the impact of policies and programs on it, and of the impact of science and technology on society and the economy.

Science and Technology indicators are constructed from various input and output statistics of the S&T system. Input indicators are mainly constructed from the financial statistics such as the level of funding for R&D, funding for basic or applied sciences. A common input S&T indicator is GERD (Gross domestic expenditure in R&D). This is the total expenditure in R&D of a country with respect to the overall expenditure. Another important input indicator is constructed from Manpower involved in R&D.

Indicators constructed from research papers, patents, standards, significant innovations, and product announcement are output indicators. They provide indication of the output and outcome of the S&T.

Common Indicators of R&D and Innovation: Strengths and Weaknesses

| Measure | Strengths | Weaknesses |
|----------------------------------|---|---|
| Financial Indicator | Helps capture how much investment is made in R&D overall by a country w.r.t. to the country's total investment. | Difficulty in identifying investment in S&T by source and by performer. Possibility of double counting |
| Level of funding in R&D activity | Captures investment discipline wise, identify priority areas. | |
| Manpower Indicator | Captures S&T personnel involved in S&T overall/ in different activities. | Difficult to capture the whole population involved. Over-estimation and under-estimation of manpower involved in different S&T activities such as teaching, R&D can happen. |
| Research Papers | Good proxy to assess scientific research. | Tacit and strategic knowledge not captured |
| Patents | Regular detailed & long term data | Uneven propensity to patent across sectors, Long complex documents. |
| Standards | Adoption indication | Standard document in any area is scattered. Difficult to properly interpret due to technical complexity. |

| Measure | Strengths | Weaknesses |
|--------------------|----------------------------|--|
| Significant | Direct measure of output | High cost of collecting the data. Difficult to delineate whether it is a significant innovation. |
| Innovations | | Misses incremental changes. |
| Innovation Surveys | Direct measure of output | What constitutes innovation can itself be questioned. |
| | Comprehensive coverage | Cost of collecting data is high. Data can suffer from reliability and validity. |
| Expert Judgments | Direct use of expertise | Finding independent expertise. |
| | | Judgments beyond expertise. |
| Product | Close commercialization to | Misses In-house process innovations. |
| Announcements | | Misses incremental product improvements. |

Financial and manpower indicators are input indicators of S&T. Indicators constructed from research papers, patents, standards, significant innovations, and product announcement are output indicators. Innovation surveys and expert surveys can capture both the input and output indications of the S&T system. Survey has data of quantitative and qualitative type. Expert has data typically of qualitative types.

3. Input Indicators

Financial indicators help to capture 'priority' of a country or units (firms, universities) to research. GERD is frequently used as an input financial indicator. GERD shows the investment in R&D of a country w.r.t. the total investment. S&T investment per capita is another indicator frequently employed to highlight S&T priority. For countries with huge populations mainly India and China this indicator will give dismal indications and may not show the real aspect one wishes to measure. The share of R&D investment in different disciplines/areas of activities; investment in basic, applied research and experimental development is used to capture research priorities of a country.

Manpower indicator: Total S&T personal of a country is applied as an indication of the scientific capacity of a country. S&T personal by their level of education further distinguishes the knowledge pool a country has. A sophisticated indicator like FTE (Full time equivalent) is used for showing actual involvement of persons in R&D activities. For example those who are involved in teaching and research though this indicator weightage is given to distinguish actual involvement in research. Say a university faculty is involved 60% of the time in teaching and 40% in research. Thus FTE of that person is 0.4. So the manpower involved in R&D of a country or units (Say University, firm) can be properly captured through this indicator. One can also

obtain indication of the demand of S&T manpower; S&T Utilisation Ratio which indicates how many S&T personal are involved in a country or units w.r.t to the total population, salary of R&D personal w.r.t. personals involved in other activities.

4. What is Scientometrics?

The quantitative approach to characterize scientific activity emerged as a new strand of research within science and technology studies in 1960's. Science becoming huge in terms of investment and skilled manpower requirement, competition for funding among different disciplines, peer review process being questioned as subjective helped push the new agenda of quantitative approach. This quantitative approach to measure scientific activity was coined as Scientometrics. It is a generic term for a system of knowledge which endeavors to study the scientific and technological system, using a variety of quantitative approaches within the area of Science and Technology Studies (STS).

Scientometrics has followed the trajectory of econometrics in the use of quantitative data, concepts and models and extensive use of mathematical and statistical technique of modelling and data analysis. Thus like economics which attempts to measure the 'health of economy', scientometrics attempts to measure the 'health of scientific and technological activity of the country, S&T institutions and S&T human resource'.

Within this quantitative approach of 'Scientometrics', a research community became very active who were largely concerned with measuring the communication process of science. This research activity is called 'bibliometrics' and largely overlaps with scientometrics and commonly one finds they are used interchangeably. Scientometrics includes both the input and output indicators whereas bibliometrics measures the output of scientific and technology activity. Bibliometric, especially evaluative bibliometrics, uses counts of publications, patents, citations and other potentially informative items to develop science and technology performance indicators.

There are implicit assumptions/propositions that underlay the utilizations and validity of bibliometric analysis.

- One of them is Activity Measurement that proposes that counts of patents and papers provide valid indicators of R&D activity in the subject areas of those patents and papers, and at the institutions from which they originate.
- The Second important proposition is Impact Measurement, in which it is proposed that the number of times those patents and papers are cited in subsequent patents or papers provides valid indicators of the impact or importance of the cited patents and papers.
- The Third important proposition is Linkage Measurement. In this it is proposed that citations from papers to papers, from patents to patents, and from patents to papers, provide indicators of intellectual linkages among the organisations that are producing the patents and papers, knowledge linkages among subject areas.

The application of Bibliometric Analysis can be under four levels: (a) Evaluation of National or Regional technical performance (policy level); (b) Evaluation of Scientific Performance of universities or technological performance of company (strategic level); (c) Tracing and Tracking R&D Activity in specific scientific and technological areas or problems (tactic level); science-technology linkage, etc. and (d) Identifying specific activities and specific people engaged in R&D (conventional level).

Elements, units and levels of Aggregation in Bibliometrics: Bibliometric Analysis is based on publications and authors; units are specific aggregates such as journals, subject categories, and institutions and countries to which papers can be assigned. References (citations) are specific elementary links between papers. When dealing with patents, inventors and assignees are relevant elements

The distinction between three levels of aggregation is important. Each level of aggregation requires its own methodological and technological approach. Micro Level: Research output of individuals and research groups; Meso Level: Research output of institutions and scientific journals; Macro level: Research output of regions and countries.

5. Scientometric Techniques

In terms of methodology, Scientometric Techniques can be classified into two categories: One-Dimensional (or scalar) and Two-Dimensional (or relational technique).

One-dimensional techniques are based on direct counts (or occurrences) and graphical representation of specific bibliometric entities (e.g., publications and patents) or particular data elements in these items, such as citations, keywords or addresses. They are used to generate scalar indicators for monitoring the S&T system. Two-Dimensional Techniques are based on co-occurrences of specific data-elements, such as co-occurrences of keywords/ classification codes, authors publishing together. The two dimensional techniques allow for capturing the network effect, relationship among entities and play an important role in understanding the thematic structure of a research field, collaboration and its impact, institutional linkages.

6. Science and Technology Performance Indicators

There are three types of matrices involved in publication based indicators:

- Publication Output Matrices: Scholarly Output, Publication Share, Publication in Top percentiles (say in Top 1% of world publication, Top 10% of world publications...), Publication in Top Journal Percentiles (top journal percentiles in terms of Impact factor);
- Citation Impact Matrices: Citation count, citation per publication, Impact factor, h index, citation share; and
- Linkage Matrices: Co-Authorship, Cross-country collaboration, Co-word Matrix.

Some Common Publication Based Indicators of Productivity are highlighted in the Table below.

| Indicator | Further Description | Advantage | Disadvantage |
|--|--|---|------------------------------|
| Numbers of papers | Based on the volume of paper produced by a country, institution, individual researchers | Easy to retrieve Gives a broad assessment of research activity | Does not inform about impact |
| Share of the number of papers | Share = (Papers from X ÷ Global output) × 100 (can also show share of different institutions in the overall publication profile of a country, research groups) | Can be useful to get relative assessment | Does not inform about impact |
| Comparison of research output over the years | International comparison of countries by “the degree of contribution to the production of papers in the world” | Evolution of research output in different years | Does not inform about impact |
| Activity in different fields | Can show the intensity of scientific activity field-wise/sub-field wise | Can be useful to see which areas are performing better if taken relative to a country/institution | Does not inform about impact |
| Co-authorship analysis | International collaboration/ National collaboration/ Department collaboration | Shows to what extent an unit cooperates with other units in the production of papers | Does not inform about impact |

Some examples to illustrate above mentioned indicators.

Example 1: Scientific publications and global share of scientific publications from India

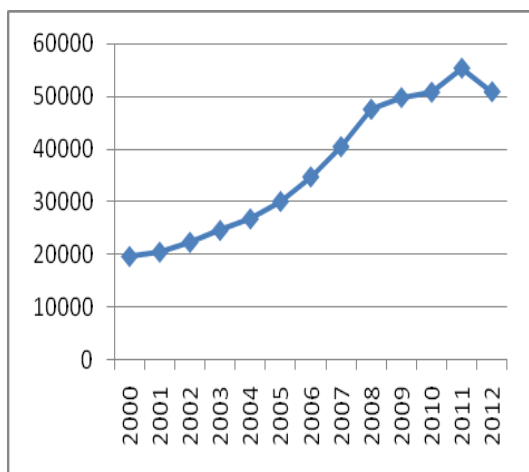


Fig 1 a) Publication output year-wise

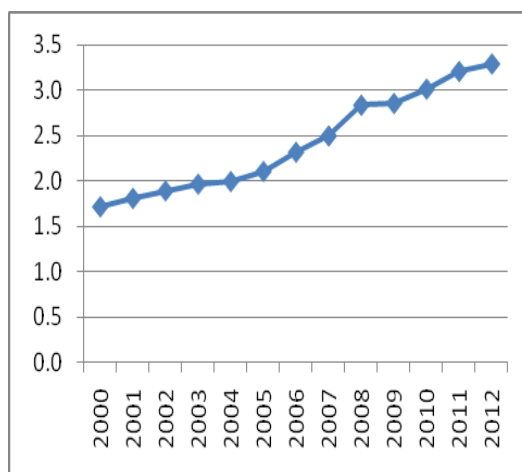
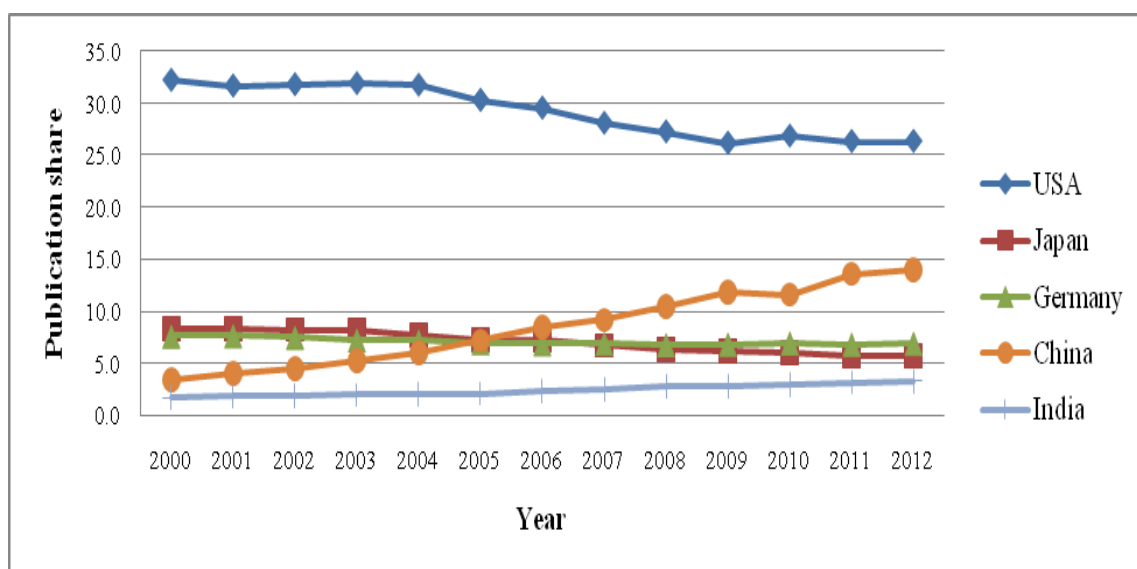


Fig 1 b) Publication share year-wise

Example 2: Share of world research output by developed countries and some emerging economies.



Example 3: Publications from India in different fields.

| Research Areas | 2000-11 | |
|--------------------------------|---------|-------|
| | Papers | Share |
| Engineering | 2424670 | 24.3 |
| Chemistry | 1621156 | 16.2 |
| Physics | 1604621 | 16.1 |
| Computer Science | 1274468 | 12.8 |
| Materials Science | 973841 | 9.7 |
| Biochemistry Molecular Biology | 916902 | 9.2 |

Example 4: Authorship pattern of Indian publication activity in nanotechnology

| Year | Single Author(Share of Publications) | Two Authors(Share of Publications) | Multi Authors(Share of Publications) |
|------|--------------------------------------|------------------------------------|--------------------------------------|
| 2000 | 13(5) | 60(24) | 173(70) |
| 2005 | 51(4) | 225(20) | 846(75) |
| 2009 | 103(3) | 718(21) | 2634(76) |

Calculating Multi-authorship share of publications– Count the number of articles published by the analysed unit during the analysed time span and check how many of them were co-authored together with a selected other unit. Divide the second figure by the first one to get the share of articles co authored between the units.

$$Px = P_x / P * 100$$

where, P_x = number of publications co-authored with the selected unit; P = total number of publications produced at the analyzed unit during the analyzed time.

For example in the above table, the total publications P is 246 (sum of single author publications, two author publications and multi-author publications). The multi-author publications P_x (M) as given is 173; percentage share of multi-authors are therefore $(173/246) * 100 = 70\%$ (rounded value)

Some Common Publication Based Indicators of Impact.

| Indicator | Further Description | Advantage | Disadvantage |
|---------------------------------|--|--|---|
| Number of citations | | Indication of a papers influence | Does not take into account that older articles usually are more cited. |
| | | | Also does not take into account that the citation rates vary between document types and subject areas |
| Citations per publication (CPP) | CPP= Total citations received/Total papers | Gives an indication of the average scientific impact | Citation rates vary between document types and subject areas |
| Citations received in the | How fast paper made impact on | Show influence of the work | Areas which are topical or addressing current |

| Indicator | Further Description | Advantage | Disadvantage |
|---|---|---|---|
| year of publication | international community | | debate have high probability of attracting immediate citations than others |
| Uncited papers | The number of papers which did not received citation even once during the time period considered | Can indicate paper is not an influential work | It can be possible that the idea is extremely novel or there are few researchers working in the subfield/topic |
| Highly cited papers a. | Number of papers that received maximum citations during the research period | Indicate paper of high value | High normalized citation score can be due to few highly-cited articles---this is not considered |
| Journal Impact Factor (IF) ^b | IF= Number of citations in year in a journal Divided by number of source items in the journal in the preceding two years. | It is assumed that high IF journals have high influence and more valuable. Thus papers in high IF journals are considered valuable. | IF is field dependent because citations have strong variance field-wise. |
| | Helps to rank journals. | | Also papers in a journal are highly skewed in citation impact (how many citations they attract) and thus IF of a journal does not truly represent paper impact in that journal. |
| Number of papers in top ranked journals | Select journals according to a suitable criterion like Impact factor of the journal | Does reflect the potential impact of paper | Does not take the size of the analyzed time duration into account |

Note: ^a For further clarity refer Example 6; For further clarity refer Example 7

Some examples to qualify the above mentioned indicators are:

Example 5: Publications from India: Nanotechnology Scenario

| Year | Publications | Citations | Citation per paper (in the year of publication) | Citations received in the year of publication (Uncited papers in the year of publication; % Uncited) | Uncited papers (% uncited)* |
|------|--------------|-----------|---|--|-----------------------------|
| 2005 | 1072 | 15985 | 14.9 (0.3) | 295 [777; 72%] | 127 (12%) |
| 2009 | 3086 | 14559 | 4.7 (0.4) | 1364 [1869;61%] | 762 (25%) |
| 2011 | 5020 | 5260 | 1.0 (0.4) | 2241 [3806;76%] | 2674 (53%) |

Example 6: Trends in Highly Cited Papers (2011)

| Country | Total Papers (rank) | Top 1% highly cited papers (rank) |
|----------|---------------------|-----------------------------------|
| USA | 455541 (1) | 9308 (1) |
| Japan | 98890 (5) | 1098 (9) |
| Germany | 118598 (3) | 2626 (2) |
| UK | 102754 (4) | 2551 (3) |
| France | 82293 (6) | 1555 (5) |
| China | 235639 (2) | 1943 (4) |
| India | 55389 (10) | 319 (20) |
| S. Korea | 53601 (11) | 533 (15) |

Note: In this example the top 1% highly cited papers in year 2011 globally are taken and the presence of different countries is shown by number of papers and their rank relatively.

Example 7: Journal Impact Factor

The 2005 impact factor of the journal Nature is produced by counting the number of citeable publications in Nature during 2005 that cite publications in Nature from 2003-2004 and dividing this with the total number of publications in Nature 2003-2004.

Description:

$$IF = \frac{C}{P}$$

where: I = the impact factor for journal J in year Y; C = the number of citations from publications in year Y to publications in journal J published Y-2 and Y-1; P = total number of citable publications in journal J in year Y-2 and Y-1.

Example 8: Publication activity in some high IF Journals in different Disciplines (Year 2012)

| Sl. No. | Journal (Impact Factor) | Total no. of publications | Share of Intl collaboration (%age of papers through intl. collaboration) |
|---------|--|---------------------------|---|
| 1 | Lancet (39.060) | 43 | 25 (58%) |
| 2 | Nature (38.597) | 20 | 13 (65%) |
| 3 | Nature reviews molecular cell biology (37.162) | 0 | 0 |
| 4 | Nature Nanotechnology (31.170) | 3 | 2 (67%) |
| 5 | Science (31.027) | 13 | 7 (54%) |
| 6 | Progress in polymer science (26.383) | 2 | 1 (50%) |
| 7 | Progress in energy and combustion science (15.089) | 2 | 1 (50%) |
| 8 | Biomaterials (7.604) | 25 | 9 (36%) |
| 9 | Water Research (4) | 4 | 1 (25%) |

The Table highlights India's publication in high IF journals are driven to a large extent by international collaboration.

7. Identifying Conceptual Connections among Documents

Indicators of conceptual linkages among papers can be constructed through matrix of co-occurrences of bibliographic units. Co-occurrence among keywords, relationship among documents based on common citations are two frequently employed methods. These indications help to show the intellectual structure of a field, research fronts and analysis undertaken over a period of time show how the intellectual domain of a field is changing. Bibliographic coupling, co-citation analysis, co-word analysis are common methods to capture these indications.

Similarity through Matching Reference (Bibliographic Coupling): A reference in an article reflects one or more concepts upon which the article draws. Two articles that share a common reference (bibliographic coupling) would therefore have some linkage through the shared concept(s), even though the articles themselves might have vastly different terminology. So, searching for linkages among two or more articles through shared references offers a way to identify linking mechanisms.

8. Co-Citation Analysis

8.1 Similarity through identifying jointly cited papers (Co-Citation)

Co-citation analysis involves tracking pairs of papers that are cited together in the source articles. When the same pairs of papers are co-cited with other papers by many authors, clusters of research begin to form. The co-cited or “core” papers in these clusters tend to share some common theme, theoretical or methodological or both.

Method: References in a document are identified. Relatedness between these references is calculated (how many times two references occurred in the same document). The references are clustered using a co-occurrence matrix. Finally, the original documents are assigned to these reference clusters

9. Co-word Analysis

Co-word analysis is a content analysis technique that uses patterns of co-occurrence of pairs of items (i.e., words or noun phrases) in texts to identify the relationships between ideas within the subject areas presented in the texts. It is used to identify the relationships between ideas within the subject areas presented in the texts and the strength of relationships between items. Co-word analysis is also very much similar to co-citation analysis. The only difference is that co-word analysis focuses on words in the document rather than references.

Method: The words or phrases that are important are identified and the relatedness between words is calculated (based on co-occurrence). Finally, the words are clustered and documents are assigned to these word clusters.

What all can be done from Publication analysis: Summary Table.

| Variables | Different Indicators which can be constructed |
|------------------|---|
| Authors | Number in a subject, field, institution, country; growth; correlation with productivity; collaboration - co-authorship, associated networks; author in a subject |
| Origin | Rates of production, size, growth by country, institution, language, subject; Correlation with economic & other indicators |
| Sources | Journals: Growth, dynamics, numbers; life cycles; quantity/yield distribution; Various distributions by subject, language, country |
| Contents | Analysis of texts -- distribution of words, phrases in various parts; subject analysis, co-word analysis |
| Citations | Citation indexes, impact factors, co-citation studies etc; Some other analysis - number of references in articles, number of citations to articles, bibliographic coupling; co-citations - author connections, subject structure, networks, maps etc; papers validation with qualitative methods and impact |

Note: Adopted from Tefko Saracevic study (from Rutgers University)

10. Methodological problems of bibliometric based indicator

Many of the problems in construction of bibliometric indicators can be addressed if one has understanding of principles behind construction of indicators. Most of indicators often have little relationship with what they Attempt to Measure? How those measurements might be carried out and used?, How the instruments that they identify influence the working of the system?

In the context of publication based indicators following limitations are primarily visible: Indicates quantity of output not quality; Non-journal methods of communication ignored; Publication practices vary across fields, journals, employing institutions; Choice of suitable, inclusive database is problematical; Undesirable publishing practices (artificially inflated number of co-authors; shorter papers); Papers represent only one output of laboratory based activity.

Citation is used as a proxy of quality but this has its own shortcomings. In particular the fact that a paper is less frequently cited or (still) unquoted several years after its publication gives information about its reception by colleagues but to what extent it indicates quality is questionable. A paper of high value may not attract citations due to variety of reasons. On the other-hand a questionable paper may attract high citations due to large number of authors questioning the results. Citations vary across field, the size of the research community among others. Lack of citation may also be due to content getting integrated into the body of knowledge of the respective subject field. Low/no citations may indicate likely that the results involved do not contribute essentially to the contemporary scientific paradigm system of the subject field in question.

Intellectual link between citing source and reference article may not always exist; Incorrect work can be highly cited; Methodological papers among most highly cited; Citations lost in automated searches due to spelling differences and inconsistencies; Similar to publication practices, citations vary across fields, journals, employing institutions; SCI and Scopus source in which citations are available changes over time; SCI and Scopus is biased over English language journals; Works of great importance rapidly become part of a common knowledge and are thus referred to in the literature without citation.

Citations may be critical rather than positive, however it has been argued that even contested results make a contribution to knowledge; The various scientific fields are cultivated by groups of varying size, and thus the probability of being cited varies from sector to sector; The number of citations does not follow a linear rate in the course of time; The value of scientific work is not always acknowledged by contemporaries.

11. Summary

This unit is designed to expose the students to the concept of indicators, the different science and technology indicators and their application. The main focus is on output indicators of S&T. The unit shows how scientometrics/bibliometrics helps to construct S&T output indicators and apply them for capturing the different facets of S&T activity including performance. Examples are given for highlighting the usage of some S&T indicators.

It is important to construct indicators that can address intersection of Input and Output indicators; for example linking funding to performance indicators. Understand limitations of indicators based on publication and citation count which can help in proper interpretation of results. This leads to wider acceptance of indicators. Tendency to make claims that are questionable should be avoided.

12. References

1. Adams, J. (2012). Collaborations: The rise of research networks. *Nature*, 490, 335–336.
2. Adams, J. (2013). Collaborations: The fourth age of research. *Nature*, 497(7451), 557–560.
3. Bhattacharya, S., Shilpa (2016). Capturing the growth dynamics of science: a publication-based analysis. *Current Science*, 110(8), 1419-1425.
4. Bhattacharya, S., Shilpa, Kaul, A. (2015). Emerging countries assertion in the global publication landscape of science: a case study of India. *Scientometrics*, 103, 387-411.
5. Elsevier, B.V. (2012). Bibliometric study of India's scientific publication outputs during 2001–2010. Study commissioned by Department of Science and Technology—NSTMIS, India.
6. Evidence. (2011). A bibliometric study of India's research output and collaboration. Study commissioned by Department of Science and Technology—NSTMIS, India. (website: http://dst.gov.in/whats_new/whats_new12/report.pdf).
7. Royal Society. (2011). *Knowledge, network and nations*. UK: Royal Society Publishing.