Scientometrics (2017) 112:697–709
DOI 10.1007/s11192-017-2334-x

Reflections on how to evaluate the professional value of scientific papers and their corresponding citations

Jaroslav Fiala[1] · Jiří J. Mareš[2] · Jaroslav Šesták[1,2]

Received: 30 January 2017 / Published online: 9 March 2017

Abstract It is inevitable that the ´publish or perish´ paradigm has implications for the quality of research published because this leads to scientific output being evaluated based on quantity and not preferably on quality. The pressure to continually publish results in the creation of predatory journals acting without quality peer review. Moreover the citation records of papers do not reflect their scientific quality but merely increase the impact of their quantity. The growth of sophisticated ´push -button´ technologies allows for easier preparation of publications while facilitating ready-to-publish data. Articles can thus be compiled merely through combining various measurements, usually without thought to their significance and to what purpose they may serve. Moreover any deep-rooted theory which contravenes mainstream assumptions is not welcomed because it challenges often long-established practice. The driving force for the production of an ever growing number of scientific papers is the need for authors to be recognised in order to be seriously considered when seeking financial support. Funding and fame are distributed to scientists according to their publication and citation scores. While the number of publications is clearly a quantitative criterion, much hope has been placed on citation analysis, which promised to serve as an adequate measure of genuine scientific value, i.e. of the quality of the scientific work.

Keywords Professional value · Citation response · Quantity versus quality · Impact factors · Databases

✉ Jaroslav Šesták
sestak@fzu.cz

Jiří J. Mareš
maresjj@fzu.cz

[1] New Technology-Research Centre in the Westbohemian Region, University of West Bohemia, Univerzitní 8, 30114 Pilsen, Czech Republic

[2] Division of Solid-State Physics, Institute of Physics, v.v.i, Academy of Sciences of the Czech Republic, Cukrovarnická 10, 16200 Prague, Czech Republic

Introduction: 'publish or perish'

The authors have been actively involved reviewing papers arising from Materials Science research. These papers form a significant portion of the 70,000 scientific publications produced annually and contribute to over one million papers (about 10 million pages) of published scientific text. Additionally, almost half of all manuscripts submitted are rejected directly by journals or as a consequence of peer-review. It is known that about half of all published papers are never read (or even noticed) and a mere one percent of publications receive over half of all citations. About 90% of all actual information is never cited, with only a few thousand scientific papers generating significant citations. In contrast to seventeenth century science where a scientist would be capable of reading almost all information published to that date, today the average scientist does not possess the capacity to even read all the papers related to their own specialization and thus is reliant on abstracts. Many scientists, moreover, do not seem read anything other than their own previously published data, and are almost irritated by information from other sources. They are focused on obtaining as many citations as possible in order to support their attempts to obtain financing.

What is the driving force behind the production of the mounting number of scientific papers? Besides the original and natural need to share and disseminate the latest knowledge, there is now an immense need to have a distinguished publication record in order to be well considered when seeking financial support. This fact was succinctly captured in the popular dictum "*Work-finish-publish*" attributed to M. Faraday (1791–1867),[1] and in the same vein, the notoriously applauded phrase: "*publish or perish*"[2] (Garfield 1996), which is worthy of serious reconsideration particularly in the age of computer facilitated production of reports.

Almost twenty years ago we published an essay (Fiala and Šesták 2000; Fiala 1987) describing the storage and citation protocols utilized in the sphere of scientific literature (Fiala 1987; Šesták 2012) noting that "*if the aim of science is the pursuit of truth, then the pursuit of information may drive people away from science*". Since then the demand for more extensive dissemination of data has accelerated because most scientific evaluations account for the ability to be *seen* (the reach of the publications), which is rated according to the journalś *impact factor* (***IF***) rating, which ranks the journal against other journals within the same field and the authorś *citation feedbacḱ* (i.e. readerś responsiveness). Specific databases have been established and available records are followed to provide a basis for a less subjective scientific appraisal, though completely objective assessment is, as yet, unreachable. The most commonly used database is the ISI '*Web of Science*' (WOS) which sets the standard for providing easily accessible data (from 1972) on a journal, paper and/or author, yielding figures based on total citation and annual citation records, as well as partial data on annual mean responsiveness (including IF and h-index). In addition, there is another larger database of peer-reviewed literature held by Elsevier—SCOPUS (Burnham 2006) which is more complex to search but is often preferred when exploring more recent data (after 1990).

In recent evaluation practice, the journal IFs are produced by the Thomson ISI Journal Citation Reports (JCR),providing a quantitative tool for journal appraisal. The *impact factor* (IF) is a measure of the frequency with which an 'average article' in a given journal

[1] http://en.wikiquote.org/wiki/Michael_Faraday.

[2] http://en.wikipedia.org/wiki/Publish_or_perish.

has been cited within an agreed period of time (often a two-year interval). It is determined by dividing the number of citations (in x) to all publications in the journal in the previous two years (N_{cit}, x) by the number of publications in this journal (N_{pub}) in the same period of time. Thus the IF is considered to be the average number of times published papers are cited up to two years after publication. Alternatively a personalized total of all citations to the authoŕs work is usually regarded as a representative indicator of his/her scientific success (i.e., authoŕs popularity). However, of greater informative value is the average citation per publication, reported as the newly introduced (by American physicist J. Hirsch at 2005) and widely accepted *h-index* which is used to measure the productivity of an individual (or group or institution), which when calculated takes into account both the number of publications and the number of citations per publication. The authoŕs h-index has a value of N, if N publications have been cited at least N times, with the remaining publications cited less than N times. For example, an author who has an h-index of 22 means that he has 22 publications which received 22 or more citations on each paper. Another indicator, the *Proceedings Citation Index* (J_{imp}) is expressed as a specific number of evaluation points for papers in international journals with IFs. The formula employed is $J_{imp} = 295 \times 10 + f$, where $f = (1 - N)/(1 + (N/0.057))$. The value of N is normalized sequence within the given category of journals, $N = (P - 1)/(P_{max} - 1)$, where P is the rank of the periodical in the field according to the descending order of the IF in the Journal Citation Report.

Publication strategies

The way in which published research is stored and made available has changed radically in recent times. Libraries have, on average, been removing about 200 outdated printed volumes each year due to shortage of space, moving to space-saving digitalization of records and computerised data-basing. Over the last twenty years there has been an ongoing debate on the best means to evaluate the ever growing amount of scientific work. This is often done based on citation analyses, specifically using the Impact Factor, and over time some problematic aspects have been revealed (Seglen 1997; Adam 2002; Scully and Lodge 2005; Lehmann et al. 2006; Editorial 2008, 2013; Frey and Rost 2010). It is surprising that in spite of the fact that some time ago there was a recommendation to abandon this method of analysis (Editorial 2003), the 2012 statement by DORA (San Francisco Declaration of Research Assessment) condemning specific international initiatives (Ylä-Herttuala 2015) remains relevant. On the other hand it should be noted that there is no doubt that citation analyses add great value as an aid in the search for scientific information.

What are the current problems with the publication process? Firstly, the sheer volume of potential publications makes assessment difficult to achieve, particularly given the variety of approaches and criteria currently used. Secondly, where high impact journals are concerned, their editors have gained a pre-eminent position in terms of deciding which papers should be admitted to the second stage of the publication process. About half of all manuscripts are rejected at this point. This editorial dominance can lead to favouritism and the operation of cliques. Even papers that contain data which can easily be disproved may still achieve publication, if the author is of sufficient standing. Conversely an author of little or no standing will often have little or no chance of having work accepted by the leading journals. Finally, any paper which challenges the existing scientific orthodoxy will also have little chance of being accepted. Even data from the most famous scientist can be unaccepted, for example the heat inertia effect of Newton's cooling law (1701) has yet to

be properly incorporated into core scientific understanding (Holba and Šesták 2015). As Max Planck (1858–1947) said, "A scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die and a new generation grows up that is familiar with it"(Wissenschaftliche Selbstbiographie (1948).

What is the consequence of this scenario? The need to publish, in order to increase the chance of securing funding, has driven a massive expansion in the number of journals available to researchers. These journals, acting in a crowded market and dependant on the income of those submitting papers for publication, often do not employ rigorous peer review or request extensive revisions in order to provide an easy path to publication. These journals often attempt to mislead the readership and appear to associate themselves with prestigious journals by using similar journal titles e.g. by using a different word order; or by using a seemingly prestigious title. They may also attempt to inflate their perceived importance by using bespoke assessment formula which favour their journal and name these measures using a variation of Impact Factor e.g. Global Impact Factor or Universal Impact Factor. Such journals are now commonly referred to as predatory journals. Fifty of these titled as the "International Journal of Advance of » Something«" even had their own Wikipedia sites providing cover for standalone predatory journals. In 2013 the shocking news appeared that predatory journals published about 420 thousand papers in contrast to the mere 53 thousand published in 2010. We can infer from this that the careers of many second-rate scientists have been supported. If this trend continues another generation of "super-predatory" periodicals may appear to fill the *´gap inside the gaṕ*.

Seemingly contradictory scientific results and the impenetrability of certain novel ideas have resulted in the formation of two scientific camps: one traditional and conservative and the other dissident, postulating ideas against the mainstream. For example, dissident alternatives to conventional physics are discussed in specialist journals such as *Apeiron* or *Galliean Thermodynamics* where challenges are often expressed as questions, such as: "Can we believe that the Maxwell equations are universally applicable?" or "Is the universe finite or infinite?" etc. Nevertheless these journals are providing a crucial platform for discussion and the development of new ideas (Fig. 1).

Fig. 1 Publishable value of scientific results or just a diplomatic or random verdict

The growth of sophisticated "push button" technologies facilitates the generation of ready to publish data and consequently simplifies the preparation of manuscripts. Articles can thus be compiled by mere combination of different measurements usually without analysis or understanding of their significance. Consequently journals overflow with elegantly written papers that describe various measurements on a mixture of materials, which are generally ignored. This situation encourages a culture which favours those skilled in formal writing and discriminates against those less skilled in this area but who focus on scientific development. In extreme situations this has led to the founding of specialist organisations dedicated to the perfection of (hopefully scientifically sound) manuscripts. *Consequently the true value of research has begun to separate from its formal valuation.* With limited financial resources university, academic, departmental and industrial workplaces are influenced, where the judgement of results now favours form over function.

Scientific information and citation (or the hunt for fame and fortune)

The purpose of a citation in a scientific publication is to facilitate the tracing of information to its source. Scientific publications (articles, journals, books, proceedings) can be indexed (recognized, retrieved) by the materials they cite. This idea was first proposed in 1955 (Garfield 1955) and implemented by Eugene Garfield in 1963 when his Institute for Scientific Information issued the first Science Citation Index (SCI) (annual edition: 9 volumes, 20,000 pages). Since then, SCI has been published annually, being continuously extended, updated and improved, making it the most effective way for readers to search for and identify relevant information.

The great success of citation indexing of scientific literature soon raised scientific institutions' administrators' interest in the application of citation analysis for performance assessment of people who produce scientific information (Johnson and Davis 1975; Roy 1976). Until the SCI, scientific activity was simply measured by the number of publications (articles, books, etc.) produced, i.e. by a purely quantitative criterion. As an analogy, chemical activity (a) was only measured by concentration (c) back in 1867 when Guldberg and Waage formulated their law. It was only in 1886 that vaṅt Hoff proposed the idea that in addition to concentration (quantity), the chemical effects of a component of a reaction also depended on what the component is like, i.e. on its quality, which in this scenario is determined by measuring partial pressures or electromotive forces (Mannchen 1965) and with this effect being expressed by the activity coefficient γ, where:

$$a = c.\gamma. \tag{1}$$

Similarly it was thought that some measure of the quality of scientific publications was needed to better assess the scientific performance of an individual. Consequently it was proposed that researchers should be evaluated not only by the number of publications but also by the rate of citation of his or her publications, because "the more frequently a publication is cited, the more it has been used by people and, therefore, the higher is its scientific contribution". This idea is believed to have first been presented by Robert K. Merton in his foreword in Garfielḋs book on citation indexing (1979). Garfield himself speaks about it in the last chapter of his book (1979), also noting that problems could arise if citation criterion of scientific work are applied in an all-inclusive, routine manner to assess quality without a more detailed analysis.

As we have noted earlier, problems have indeed arisen. The book "The Web of Knowledge" published to honour Eugen Garfield on the occasion of his 80th birthday (Cronin and Atkins 2000) discusses this in 151 out of the total 565 pages. It showed that the mere rate of citation of a publication was not, in many cases, an adequate measure of its scientific value. A number of modifications were recommended and more sophisticated criteria developed, taking into account other aspects in addition to the rate of citation, such as the h-index (Hirsch 2005), g-index (Eghe 2006) and others (Bornmann et al. 2011), but ultimately this has not yet improved the situation to any great extent.

Recognition and legitimacy

What is behind the failed attempts to *measure science*? Is it the idea that the main thing that motivates researchers in science is that money and honours are distributed on the basis of the results of such measurement? This idea is not only a tragic mistake, but is also the main reason behind the dissatisfaction of the scientific community with the manner in which its activities are currently assessed. Furthermore the efficiency of citation indexing as a tool for retrieving scientific information has decreased greatly since the time citation started to be used to assess scientific work. This is because information, i.e. communicable knowledge, while the principal product of science, is not its ultimate goal. After all, the goal of science is to know (i.e. to learn and to recognize) the truth. As quoting J.W. Goethe wrote (1870):

> that I may detect the inmost force
> which binds the world, and guides its course;
> its germs, productive powers explore,
> and rummage in empty words no more!

Citations of publications are of great help in improving the efficiency of retrieving information (identification) but are completely unsuitable for the assessment of individual scientists. One reason for this is, due to the abundance of scientists and scientific work, so the methods used in assessment are necessarily of statistical nature and are relevant for large sets only. Citation analysis can be used to compare the scientific activities of (approximately equally) big countries, giant scientific institutions, or magazines that publish a great number of articles year on year. Even in these cases great care is needed because statistically valid results are often only produced over time spans which are longer than the entire careers of individual scientists (Ketcham 2007; Garfield 1999). Material support and social recognition of individuals are based on the evaluation of their scientific activities. But the resources of both are limited, which tends to encourage collusion between the citing and the cited subjects, creating closed groups who have the aim of establishing a dominant position in procuring funds and fame. This is probably what the King of Bohemia and Holy Roman Emperor Charles IV had in mind when they established in the Founding Charter of Prague (Charleś) University in 1348 that it was the duty of students to swear that they would continue to devote themselves to science after leaving the University "not for filthy profit or passing fame but to propagate the truth and to brighten its light upon which the human welfare rests". Money and fame are temporary worldly possessions but the truth is eternal and transcendent, really the "light of the world"…Goethe said, as he was dying "*Mehr Licht!*", and Jesus said "I am the light of the world".[3] Moral weakness leads humankind to strive for personal profit and fame far more often than to strive to uncover universal truths.

[3] The Gospel according to St.John 8:12.

Evaluation of scientists based on citations only strengthens this impulse and creates conflict within the scientific community. As "every kingdom divided against itself is brought to desolation",[4] evaluation of scientists based on citation becomes the driving force for decline of the scientific community. "There will be no hope for honesty, peace and abundance in the world should individuals only pursue their own interests according to their ideas and should they not be attracted to and connected with something in common, to understand clearly that only together they can achieve success in everything" by the Bohemian thinker Jan Amos Komenský (1592–1670). If scientists are evaluated by citations, it means that (owing to the cartelization of the scientific community), in the final analysis they only evaluate themselves. But for such an evaluation to be fair, it is beyond their control: "it is not in man that walketh to direct his steps".[5]

Basis of misunderstanding

Throughout the history of science time and money have been two of the major factors limiting research. In England it used to be said that if a man wanted to do science, he had to have a large manor and a reliable administrator to send him one thousand pounds every fortnight. It was in this way that Joseph Achille le Bel (1847–1930) was able to practice science and was consequently able, together with vaṅt Hoff, to propose the idea that the four bonds of carbon are not oriented randomly, but have a specific spatial arrangement. Thanks to the wealth of his family, he did not have to work to earn his living and could set up his own private laboratory. Likewise John William Strutt, Baron Raleigh (1842–1919), who inherited an estate with seven thousands acres of land (which his younger brother agreed to manage) was able to engage himself in the theory of wave motion and its applications in acoustics, optics, and electromagnetism (Nobel Prize for physics in 1904). If someone who had no such proverbial large manor wanted to do science, he had to earn his living and carry out science as a hobby. He was the sponsor of his own scientific work, which he performed with love and pleasure as the best way of making its results useful for mankind. Others such as Johannes Kepler (1571–1630), secured a wealthy patron. Kepler made his living as an astrologer for the Holy Roman Emperor Rudolph II and was therefore able to carry out a great deal of scientific work in astronomy and optics and produced the essay "*Strena seu de nive sexangula*" (1612), considered by the International Union of Crystallography to be the very first scientific monograph in crystallography. Even when funds were available researchers needed to overcome other hurdles. Many women, such as Marie Skłodowska-Curie (1867–1934) double Nobel Prize laureate (1903 and 1911), or Dorothy Mary Hodgkin (1910–1994) Nobel Prize laureate for the x-ray diffraction research of structures of biologically important substances (1964), conducted excellent scientific work while coping with the additional duties of raising their children and taking care of their husbands (Kraus 2015).

However, great research is not only dependant on money and time. Where researchers were professionally employed, their most significant discoveries were not always the intended focus of their research. An interesting illustration of this and of researchers' attitude to their discoveries is shown by the contrasting actions of Edison and Roentgen. The laureate of the first Nobel Prize for physics, Wilhelm Conrad Roentgen (1845–1923), worked as a professor and an academic worker at the universities of Würzburg and Munich when he discovered "X-rays", still sometimes referred to as Roentgen radiation in

[4] The Gospel according to St.Matthew 12:25.

[5] The Book of the prophet Jeremiah 10:23.

remembrance of the discovery. However Roentgen was not supported for the purpose of discovering X-rays, in fact no one had anticipated that anything like X-rays existed. In the afternoon of 8 December 1895, not even Roentgen himself had the faintest idea that he would discover X-rays later that evening in his apartment. His attitude to his discovery was unusual in that he even refused to have the discovery patented. When prompted to do so by Thomas Alva Edison, Roentgen said that he would have felt ashamed should he assume even a small part of what X-rays could bring to mankind. He donated the money he received together with the Nobel Prize to the University of Würzburg. He was ascetically modest. He was not interested in any honorary degrees or functions and he did not even accept the Order of the Crown through which Prince Luitpold of Bavaria promoted him to nobility in 1896 (Kraus 1997). Unlike him, Thomas Alva Edison (1847–1931) died a rich man. For his research activities (the results of which could be envied by many renowned scientists), this talented and unprecedentedly hard-working man did not need any grant support, earning his living as a private person in his laboratories in Menlo Park, New Jersey. With his lifés credo—four hours of daily sleep is a need, five hours means leisure and six hours means laziness—he could not use the money he earned through his work on anything other than to continue to support his work and further discoveries. That scientific productivity does not need to be supported with social honours was proved for example, by Josiah Willard Gibbs (1839—1903). This man, whom Albert Einstein appraised as one of the most original thinkers-scientists the United States gave mankind, was the father of vector analysis, co-author of statistical physics and pioneer in the physical chemistry of interfaces. Yet he never aspired to membership in scientific institutions and, in the United States, his work became recognised as late as two years after publication.

In the distant past Aristoteles recognized that human beings have an innate spontaneous interest to learn the truth. This interest is the intrinsic driver of science, needs to be nurtured. However care needs to be taken to stimulate scientific endeavour in the correct way as there is a danger that the pursuit of scientific truth could be disrupted by the wrong kind of stimulus and not taking individual's characteristics into account. A physical analogy would be in the field of magnetism. Copper, silver and gold are only slightly magnetic substances; when placed in a magnetic field, small magnetic moments are induced in their atoms in a direction opposite to that of the external magnetic field, which becomes slightly weakened. In contrast, at room temperature, iron is a strongly magnetic (ferromagnetic) metal. Its atoms possess large magnetic moments. They are so strong that, by their mutual interaction, they are oriented spontaneously into Weiss domains even outside the external magnetic field. When placed in an external magnetic field, they turn to align with its direction, increasing the field considerably. However, if iron is heated up above the Curie temperature, the heat will disturb the arrangement of the magnetic moments of its atoms, iron will stop being ferromagnetic and its ability to increase the external magnetic field will be reduced by several orders of magnitude. Likewise, in some cases, money and fame allocated by authorities on the basis of problematic scientometric criteria can actually, for many of us, suppress the interest to seek scientific truth.

Quality times quantity is constant

The number of scientific publications is rapidly increasing year on year, this does not necessarily mean that the amount of in depth knowledge is increasing at the same rate. This can be illustrated by the rate of growth of the number of substances registered in the Registry File (RF) of the Chemical Abstracts Service (CAS). During its first year (1965),

211,934 substances were registered. During the year 2000, 6031,378 substances, including 5,131,250 biosequences (aminoacids in proteins and bases in nuclear acids) were registered, so that by the 31st of December 2000 at total of, 28,499,942 substances including 10,938,676 biosequences were registered in the RF. The Abstracts File—AF, CAS' second largest database, which has been in existence since 1907, increased in the course of the year 2000 by 725,195 abstracts, with the total number of abstracts in AF reaching 19,754,207 by the end of 2000. The growth in the number of registered substances from 1965 till 2000 is shown in Table 1. Since 2000, the number of registered substances has been increasing even more swiftly. So, e.g., 41,911,919 organic and inorganic substances and 60,642,927 biosequences were registered in CAS RF at 04:57:16 EST in 19.1.2009. The number of substances registered during the last six years is given in Table 2.

Unfortunately, the number of registered substances ($\approx$100,000,000 substances) is much greater than the number of substances for which we have data on their molecular structure, i.e. x–y–z coordinates of the atoms in the molecule. In fact we know this for no more than 1,000,000 substances and the difference between these two numbers is rapidly increasing. This is important e.g. in pharmacy, as knowledge of the molecular structure is required for each new medicine along with the successful passing of rigorous tests, the cost of which is estimated to be about 1,000,000,000 $ per new product. In the year 2000 some 160 new medicines were successfully developed, while only 10 new medicines arose in 2012. For the majority of substances registered in CAS RF we do not know much about their properties, how they react with other substances and to what purpose they could serve. Detailed information on inorganic and metallorganic substances used to be systematically gathered and extensively published by the Gmelin Institute for Inorganic Chemistry in the Gmelin Handbook of Inorganic Chemistry—first edition (in German) in 1817. The Institute

Table 1 Growth in the number of substances registered in the database Chemical Abstracts Service Registry File

Year	Δ	Σ	Year	Δ	Σ
1965	211,934	211,934	1983	418,905	6,346,713
1966	313,763	525,697	1984	563,390	6,910,103
1967	270,782	796,479	1985	544,618	7,454,721
1968	230,321	1,026,800	1986	628,966	8,083,687
1969	287,048	1,313,848	1987	610,480	8,694,167
1970	288,085	1,601,933	1988	602,465	9,296,632
1971	351,514	1,953,447	1989	615,987	9,912,619
1972	277,563	2,231,010	1990	663,342	10,575,961
1973	437,202	2,668,212	1991	684,252	11,260,213
1974	319,808	2,988,020	1992	690,313	11,950,526
1975	372,492	3,360,512	1993	680,230	12,630,756
1976	347,515	3,708,027	1994	777,212	13,407,968
1977	369,676	4,077,703	1995	1,186,334	14,594,302
1978	364,226	4,441,929	1996	1,269,246	15,863,548
1979	346,062	4,787,991	1997	1,376,942	17,240,490
1980	353,881	5,141,872	1998	1,679,913	18,920,403
1981	424,230	5,566,102	1999	3,548,161	22,468,564
1982	361,706	5,927,808	2000	6,031,378	28,499,942

Table 2 Total number of organic and inorganic substances (without biosequences) registered in the last years (counts in December)

Year	Count
2010	56,259,436
2011	64,765,463
2012	70,082,806
2013	76,705,260
2014	91,022,519
2015	104,228,986

was, in the final years of its existence, staffed by about 120 full-time employees, of whom about 80 had doctorates. The Gmelin Handbook (GH) presented many valuable tables of numerical data, curves, and other graphic material, including diagrams of apparatus. The GH reported the applied or "practical aspects" of the molecules and methods of their manufacture. It included about 20% of patents considered. Due to the detailed, in-depth processing of the work, data could be delayed from 2 to 25 years before appearing in new volumes of the GH. Altogether 760 volumes of the GH were issued (comprising more than 240,000 pages), plus the "Gmelin Formula Index" in 35 volumes. In the year 1997 publication of the GH was stopped and the Gmelin Institute dissolved. In addition, the basic reference work (database) on organic compounds was the Beilstein Handbook of Organic Chemistry, which was published by the Beilsten Institute for Organic Chemistry in Frankfurt-am-Main. In the final years of its existence, the Institute had 160 full-time employees, of whom 110 had a doctorate in chemistry, and more than 500 outside contributors. The Beilsten Handbook (BH) reported structural diagrams, information on natural occurrence, techniques for isolation from natural products, methods of preparation and manufacture, physical properties alone and in mixture with other compounds, chemical properties, methods of analytical determination as well as data on salts and additional compounds. The time lag between the publication of original data and the publication in the pertinent volume of BH was about 20 years. Altogether 503 volumes of BH were issued with 440,814 pages. In the year 1998 the publishing of BH was stopped.

Numbers and information

The number of substances registered by CAS RF during the past 50 years ($m = 100{,}000{,}000$) is immense. However, in practice we mostly use mixtures of substances rather than pure substances and it is as important to know how the substances behave in combination, how they react with each other. We therefore need information on $m \cdot (m - 1) = 10^{16}$ binary mixtures and on $m \cdot (m - 1) \cdot (m - 2) = 10^{24}$ mixtures. More useful, from scientific, technological and economical points of view, would be to provide information on binary mixtures of 10,000 substances or on ternary mixtures of 470 substances than on those (pure) 100,000,000 substances which are registered today. It would be better to know more about a small number of substances than little about a great number.

Among the large number of registered substances there are certainly quite a few that, in mixture with other substances, would prove to be remarkable catalysts, medicines, explosives or maybe structural materials. A major problem is that for 99% of the 100,000,000 registered substances there are no reference spectra of identification features, with the help of which those substances could be recognized. The worlds largest database

for identification of substances—Powder Diffraction File of the International Centre for Diffraction Data—contains at present only about 800,000 x-ray diffraction reference spectra. Along with the continued registration of new substances we should also register the reference spectra of all substances that have already been registered; the spectra which could help those substances to, at least theoretically, be identified. But even in the case that such a database would exist, the identification of substances registered in CAS RF will face serious problems. Expressing the reference spectra (IR, XRD, MS, NMR,…) by n-dimensional pattern vectors (Fiala 1972, 1976, 1980, 1982; Fiala and Říha 2014; Malinowski and Howery 1980; Alves et al. 2016; Qian et al. 2017), the identification of an unknown substance (unknown mixture of substances) means finding m numbers $c_1, c_2, \ldots, c_m$, which minimize the value of the residual misfit

$$\left| \vec{x} - \sum_{j=1}^{m} c_j \vec{y}_j \right| \tag{2}$$

where $\vec{x} = \left[y_{j1}, y_{j2}, \ldots, y_{jm}\right]$ is the pattern vector (vectorial representation of the spectrum of identification features) of the unknown substance and $\vec{y} = \left[y_{j1}, y_{j2}, \ldots, y_{jm}\right]$; $j = 1, 2, \ldots, m$ are reference pattern vectors. Among the m reference spectra (pattern vectors), there are only n independent vectors and each other pattern vector can be expressed as a linear superposition of those n independent pattern vectors. After the identification technique used, n amounts to several hundred or at most a few thousand. This is why the identification of $m = 100{,}000{,}000$ substances would be hopelessly ambitious, unless the identified substance is composed of only several (not more than four or five) components. In this case, the spectrum of such a simple mixture is similar to spectra of each of its components, so only comparing with those reference spectra that are most similar to the unknown spectrum will work. The efficiency of such a procedure can be increased using factor analysis. The analyzed mixture (with the pattern vector $\vec{x}$) is separated into $p - 1$ fractions and their spectra $\vec{x} \equiv \vec{x}, \vec{x}_2, \ldots, \vec{x}$ are determined. Designating the spectra of (unknown) components of the analysed substance, $\vec{z}_1, \vec{z}_2, \ldots, \vec{z}_k$, then the vectors form a k-dimensional subspace (of the n-dimensional space of pattern vectors) with the base $\vec{z}_1, \vec{z}_2, \ldots, \vec{z}_k$, which can be calculated (reconstructed) by factor analysis from the p ($\geq$k) vectors $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_p$ of this subspace (Malinowski and Howery 1980).

$$\vec{x}_i = \sum_{j=1}^{k} c_{ij} \vec{z}_j; \quad i = 1, 2, \ldots, p \tag{3}$$

If the spectra of "pure" substances (single component "mixtures") are recognized, then their identification would be straightforward when using a database $\{\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_m\}$ of reference spectra of known substances. Of course, thus far we have at our disposal a database of reference spectra for all substances which we want to identify to date. If we do not want to identify the substances, then perhaps it is needless for them to be registered.

Acknowledgement The present work was supported by Institutional Research Plan of Institute of Physics ASCR,v.v.i., and developed at its Join Research Laboratory with the New Technologies Centre of the University of West Bohemia in Pilzen (the CENTEM project, reg. no. CZ.1.05/2.1.00/03.0088 that is co-funded from the ERDF as a part of the MEYS—Ministry of Education, Youth and Sports OP RDI Program and, in the follow-up sustainability stage supported through the CENTEM PLUS LO 1402). The paper is based on a long lasting earnest letter-friendship of Jaroslav Fiala with Eugene Garfield (Institute for Scientific Information, USA). Kind interest by Wolfgang Glänzel (University of Leuven, Belgium) is appreciated.

References

Adam, D. (2002). Citation analysis: The counting house. *Nature, 415,* 726–729.

Alves, A. D., Yanasse, H. H., & Soma, N. Y. (2016). An analysis of bibliometric indicators to JCR according to Benford's law. *Scientometrics, 107,* 1489–1499.

Bornmann, L., Mutz, R., Hug, S. E., & Daniel, H.-D. (2011). A multilevel meta-analysis of studies reporting correlations between the h-index and 37 different h-index variants. *Informetrics, 5,* 346–359.

Burnham, J. F. (2006). SCOPUS database: A review. *Biomedical Digital Libraries, 3,* 1–7.

Cronin, B., & Atkins, H. B. (Eds.). (2000). *The web of knowledge, Information Today*. New Jersey: Medford.

Editorial. (2003). Deciphering impact factors. *Nature Neuroscience*, *6*, 783.

Editorial. (2008). Papers about papers, *Nature Nanotechnology*, *3*, 633.

Editorial. (2013). Beware the impact factor. *Nature Materials*, *12*, 89–91.

Eghe, L. (2006). Theory and practise of the g-index. *Scientometrics, 69,* 131–152.

Fiala, J. (1972). Algebraic conception of the powder diffraction identification system. *Journal of Physics D: Applied Physics, 5,* 1874–1876.

Fiala, J. (1976). Optimization of powder-diffraction identification. *Journal of Applied Crystallography, 9,* 429–432.

Fiala, J. (1980). Powder diffraction analysis of a three-component sample. *Analytical Chemistry, 52,* 1300–1304.

Fiala, J. (1982). A new method for powder diffraction phase analysis. *Crystal Research and Technology, 17,* 643–650.

Fiala, J. (1987). Information flood: fiction and reality. *Thermochimica Acta, 110,* 11–22.

Fiala, J., & Říha, J. (2014). X-ray diffraction analysis of materials. *Hutnické listy, 67,* 2–7.

Fiala, J., & Šesták, J. (2000). Databases in material science: Contemporary state and future. *J Thermal Anal Calor, 60,* 1101–1110.

Frey, B. S., & Rost, K. (2010). Do rankings reflect research quality? *Journal of Applied Ecology, 13,* 1–38.

Garfield, E. (1955). A new dimension in documentation through association of ideas. *Science, 122,* 108–111.

Garfield, E. (1979a). *Citation indexing*. New York: Wiley.

Garfield, E. (1979b). Perspective on citation analysis of scientists, Chapter 10. In E. Garfield (Ed.), *Citation indexing*. New York: Wiley.

Garfield, E. (1996). What is the primordial reference for the phrase´Publish or perish? *Scientist, 10,* 11–17.

Garfield, E. (1999). Journal impact factor: A brief review. *Canadian Medical Association Journal, 161,* 979–980.

Goethe, J. W. (1870). *Faust a tragedy, translated by Bayard Taylor, part I, scene I. Night*. Boston, New York: Houghton Mifflin Company.

Hirsch, J. E. (2005). An index to quantify an individuaĺs scientific research output. *Proceedings of the National Academy of the Science USA, 102,* 16569–16572.

Holba, P., & Šesták, J. (2015). Heat inertia and its role in thermal analysis. *Journal of Thermal Analysis and Calorimetry, 121,* 303–307.

Johnson, A. A., & Davis, R. B. (1975). The research productivity of academic materials scientists. *Journal of Metals., 27,* 28–29.

Ketcham, C. M. (2007). Predicting impact factor one year in advance. *Laboratory Investigation, 87,* 520–526.

Kraus, I. (1997). *Wilhelm Conrad Röntgen, dědic št'astné náhody (Wilhelm Conrad Röntgen: The heritage of lucky coincidence)*. Praha: Prometheus.

Kraus, I. (2015). Ženy v dějinách matematiky, fyziky a astronomie (Ladies in the history of mathematics and physics), Česká technika—nakladatelství ČVUT, Praha.

Lehmann, S., Jackson, A. D., & Lautrup, B. E. (2006). Measures for measures. *Nature, 444,* 1003–1004.

Malinowski, E. R., & Howery, D. G. (1980). *Factor analysis in chemistry*. New York: Wiley.

Mannchen, W. (1965). *Einführung in die Thermodynamik der Mischphasen*. Leipzig: VEB Deutscher Verlag für Grundstoffindustrie.

Newton, I. (1701). Scale graduum caloris. Calorum descriptiones & signa. *Philosophical Transactions, 22,* 824–829.

Qian, Y., Rong, W., Jiang, N., Tang, J., & Xiong, Z. (2017). Citation regression analysis of computer science publications in different ranking categories and subfields. *Scientometrics, 108,* 1–24. doi:10.1007/s11192-016-2235-4.

Roy, R. (1976). Comments on citation study of materials science departments. *Journal of Metals, 28,* 29–30.

Scully, C., & Lodge, H. (2005). Impact factors and their significance; overrated or misused? *British Dental Journal, 198,* 391–393.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *British Medical Journal, 314,* 498–502.

Šesták, J. (2012). Citation records and some forgotten anniversaries in the field of thermal analysis. *Journal of Thermal Analysis Calorimetry, 108,* 511–518.

Wissenschaftliche Selbstbiographie. Mit einem Bildnis und der von Max von Laue gehaltenen Traueransprache. Johann Ambrosius Barth Verlag (Leipzig 1948), p. 22, as translated in Scientific Autobiography and Other Papers, trans. F. Gaynor (New York, 1949), pp. 33–34 (as cited in T. S. Kuhn, The Structure of Scientific Revolutions).

Ylä-Herttuala, S. (2015). From the impact factor to DORA and the scientific content of articles. *Molecular Therapy, 23,* 609.