Nicola De Bellis

Bibliometrics and research evaluation

What's in it for the librarian?

Research evaluation in European universities has a long history of fatal attraction to bibliometrics, most notably citation analysis, by virtue of its supposed ability to provide objective indicators of scientific quality. Extensive bibliometric analyses involving collective entities (research groups, institutions, nations) are usually carried out either by scientometric scholars in the process of validating their theoretical constructs or by trained personnel during national research assessment exercises. When it comes to the evaluation of individual performance for academic promotion or funding, however, researchers willing to supply bibliometric evidence of their impact are on their own unless they turn to the local library for assistance. The present paper will 1) discuss the limitations and biases implicit in any bibliometric-based evaluation exercise at the individual level within the current framework; 2) outline a possible escape route by introducing contextual information relative to the empirical, methodological and epistemological background of the individual author's bibliometric profile.

Over the last decade, bibliometrics has become a regular in academic libraries due to the ever increasing demand for quantitative data on output and citation performance retrieved from bibliographic databases. The purpose of this brief communication is to discuss the role librarians can play in a bibliometric-informed research assessment system at the micro-level of the individual scientist willing to provide bibliometric evidence of his or her performance. The main question is: Can librarians go beyond the retrieval part of the job? Can they contribute to the construction of a better evaluation environment? If so, how?

Let's start with what is already known for sure: the standard definition of bibliometrics as

> „the application of mathematics and statistical methods to books and other media of communication" (Pritchard 1969, 349).

Just like many similar definitions of bibliometrics and related fields (scientometrics, informetrics), this one is neutral, aseptic, strongly biased toward the methodological aspect of the discipline. Basically, it reduces bibliometrics to applied mathematics. This is, by contrast, my definition:

> (Biblio/Sciento/Infor)metrics: a social science with a clearly identifiable object – the statistical regularities pertaining to the production and flow of any kind of (scientific) information solidified in some kind of documents in whatever form – which doesn't require for the documents to be read in order to support general conclusions on such quality issues as the relevance, utility, impact, ultimately contribution to knowledge advancement of their content and authors.

To claim that bibliometrics is a social science amounts to saying that it is not just applied mathematics. It deals with real people engaged in real life situa-

tions, most of which have nothing to do with pure science or applied mathematics. Besides, the objects of a social science are not natural objects existing out there independently of the observer. They are theoretical constructs that interact with the reality they are meant to represent (for example, the citation behavior of scientists can't be safely assumed to be the same now as it was before the invention and the legitimation of the political use of the *Science Citation Index*).

To state that bibliometrics deals with documents amounts to saying that its grasp of the scientific research process is quite poor because of its exclusive focus on the final product: the article, the book, the report. But what is found in the final product might not tell the whole story and, if post-Popperian epistemologists get it right, it might not even tell a true one. There is much more to science than the final report. There are many invisible, uncountable, unpredictable dark sides of scientific communication that should be taken into account for an objective evaluation of the published results.

By referring to „statistical regularities", the definition points out that bibliometrics is built on statistical analysis. It is not just counting and ranking. It is – it should be – counting and ranking coupled with measuring the uncertainty inherent to the results.

Finally, to declare that bibliometrics pinpoints quality without getting its hands dirty in the muddy waters of textual interpretation highlights the most distinguishing feature of its attitude toward science, namely, the constant, unabated tendency to transubstantiate quantity into quality: since high citation scores are usually correlated with favourable peer reviewing, more citations = more impact = more usefulness = more quality = more ... whatever. That is what I call the bibliometric leap. The leap from quantity to quality has been possible after the invention and large-scale diffusion of the tool that made citation analysis possible during the 1960s: Eugene Garfield's *Science Citation Index*. Moreover, such a leap requires the commitment to a peculiar theory of the citation process that can be traced back to the American sociologist Robert Merton: The act of citing, in Merton's view, is the same as the act of peer reviewing, only on a smaller scale. Citations are atoms of peer reviewing. Hence more citations amount to more endorsements by qualified peers.

Under certain conditions, the leap is justified and this is what keeps the bibliometric machine running. Leading bibliometric schools, such as the Leiden School and the Hungarian School, typically work with large samples of publications and citations at the level of the research group, the department, the university, the country, and they build relative, normalized indicators of performance based on standard or reference values defined at the journal level. In doing so, they account for the differences in citation cultures across disciplines and research areas, and they overcome the obstacle arising at the lowest levels of aggregation, the no-man's land of the individual scientist. They just

get around the obstacle though, they don't remove it. Indeed, when the citation scores of two scholars working in different fields have to be compared, the trick of normalizing those scores against a baseline, expected value established with reference to journal-dependent subject categories might not work. Journals can be taken as field-delimiting devices only in a very rough way and even if their classificatory function is not questioned, to compare individual scientists on account of the ideal number of times their papers could have been cited is a very poor shortcut to quality judgements: what about the skewing effect of highly cited papers (the stuff bibliometric dreams are made of) on those expected values? And what about the comparison of normalized figures only marginally different: Does the "quality" of the corresponding performers differ according to the same ratio-scale, the zero point being tantamount to zero quality?

Bibliometrics has been terrified by the individual since the very beginning:

> „[IN BIBLIOMETRICS] The method to be used is similar to that of thermodynamics [...] One does not fix one's gaze on a specific molecule called George, traveling at a specific velocity and being in a specific place at some given instant; one considers only an average of the total assemblage in which some molecules are faster than others ..." (Price 1986, xvi).

Back then, it must be said, Derek John de Solla Price's project was truly revolutionary. It disclosed a new paradigm for science studies, but it also opened up a Pandora's box of unanswered questions, the most crucial one being the bibliometric status of the molecule „George" ruled out by the thermodynamic model. That George „traveling at a specific velocity and being in a specific place at some given instant" is indeed the subject of many current bibliometric exercises. He is the true main character of many stories unfolding every single day in academic audits all around the world. And it is exactly George's unique profile that makes visible the basic message I wish to convey here: for bibliometrics to be useful in research evaluation, the starting point cannot dwell in the thermodynamic stronghold. The ambiguities and risks of micro-level analyses have to be tackled as well and we, the librarians, are in a good position to intervene because who is George, anyway? George is nothing else than the individual scholar that most likely will turn to the local library for help in the calculation of bibliometric indicators of scientific performance, like the citation score and the h-index, for a variety of purposes, ranging from academic recruitment or promotion to grant application.

In this scenario, there are at least three serious obstacles to any individual bibliometric analysis: an empirical gap, an epistemological gap, and a methodological gap.

1) Empirical gap

There is first of all a problem of empirical base. Let's pretend for a while that citations, individually taken, are a reliable evidence of impact/utility/influence, which they are not, but let's pretend they are. Then we have to col-

lect this kind of evidence wherever it can be collected and, as is well known, since 2004 the noblest descendant in the line of citation indexes, Thomson Reuter's Web of Science (WoS), in not alone anymore in the interdisciplinary counting game: there is Scopus, there is Google Scholar (GS) and, for the sake of plurality, there are also several disciplinary databases that allow citation counting in well-circumscribed fields of inquiry, such as physics, chemistry, computer science. Evidence exists that in certain domains of the natural and biomedical sciences and for relatively recent time spans, WoS and Scopus provide similar, quite comparable results, but also evidence exists that GS is unique in retrieving unique citations in several disciplines. Plus, when it comes to the social sciences and the humanities, neither of the two subscription-based services seems to do such a good job as GS, with all its limitations.

2) Epistemological gap

The entire building of bibliometrics rests, from a philosophical point of view, on Merton's normative theory of science. Here scientific research can be pictured as a staircase where each step corresponds to a well defined, clearly identifiable stage of communication and reward. And each step is bound to the following step in a seamless fashion. First come the mental and experimental operations leading to the formulation and testing of scientific hypotheses. Then, results and their interpretations are communicated to the audience of the scientists working in the same domain. They evaluate the results both directly through the peer reviewing process and indirectly by citing them in their publications. If it turns out that the results are good, then the stage is set for higher, more concrete forms of reward, such as career advancement, honors, prizes, and so on. Unfortunately, though, this is a rather mythological view of the way science works. If we just limit our attention to the evaluation part, the equivalence between citations and peer review cannot be taken for granted at the micro-level. Bibliographic citations are not always nor they need to be atoms of peer recognition. There are too many non-quality related reasons for citing and being cited.

3) Methodological gap

Individual researchers, especially if they are at the beginning of an academic career, do not usually have had the time to publish many articles and get a substantial number of citations. Maybe they have worked for a long time on a research project leading to just one or two multi-authored publications. Unfortunately, small samples are not appropriate for standard statistical analyses and comparisons. And even when the samples are not so small, in the sense that the global output and citation score of the candidate appear remarkably high, the distribution of values is typically skewed: there will be just one or two publications playing the lion's share in George's curriculum against a great many number of poorly cited or un-cited ones. Consequently, comparisons based on average values can be misleading. On top of all this, the most popular bibliometric indexes, notably the h-index and all its subsequent cor-

rections, modifications and normalizations, are increasingly being criticized by top bibliometricians both for their inconsistencies and because they flatten the complexity of the research performance by not taking into account the many factors influencing the citation process.

So what can we do, as librarians, to fill the gaps, what is the librarian's antidote? We can't compete with professional bibliometricians in the construction of new indicators, that is not our job to do. On the other hand, we can supply complementary information to help place those indicators in a richer, more (story)-telling environment. In simplest terms, we can provide contexts.

Anti-1) Anti-empirical gap

At the empirical level, providing contexts amounts to three operations. First, within a single database, the citation search should be as accurate as possible. In WoS and Scopus, for instance, we can't just take for granted the figures displayed in the citation report produced automatically by the system because those figures refer only to source items, i.e. items published in the journals covered by the database. They don't take into account orphan and stray citations, namely, citations pointing to documents that are not included among the sources. To retrieve these lost citations we have to run a genuine citation search and then add manually the lost citations to the ones retrieved in the first place. Second, since each database covers a distinct set of sources and since we are not working at the thermodynamic level where small numbers can be neglected, the search should be run across at least all the three major interdisciplinary citation indexes plus one disciplinary index. In a citation report of a chemist, for instance, the CAS databases are likely to supply the analyst with unique citations that don't show up in WoS, Scopus, GS. Third, the results obtained should be combined in one single report merging duplicate entries in order to obtain an ideal „true citation score". For now, a job like this has to done manually, but I don't see insuperable technical difficulties in handling these operations by means of an ad-hoc computer software.

Anti 2) Anti-epistemological gap

The anti-epistemological gap can be addressed only recognizing that, at the micro-level, all citations are (counted as) equal, but some citations are more equal than others. To begin with, a true self-citation rate should be calculated, which is not that easy even in WoS. More importantly, not all citations have the same meaning. Sometimes the cited documents have been actually used by the citing authors. In other cases the references are unsubstantial, simply perfunctory or ceremonial. So the context of at least a sample of citations should be addressed to understand the role played by the cited documents in the citing source. Of course, we can't analyze the context of all the citations to all the articles published by an author. We need to work on a sample. But the sample does not have to be random, quite the opposite, it has to be a carefully selected sample. As a matter of fact, in bibliometrics we can't just apply stan-

dard statistical sampling procedures because a random sample could leave out the best part, notably the most influential articles. So the reverse procedure is the good one: first we select, with the help of the author, the best candidate documents, the ones that most likely have had an impact on the citing publications, then we retrieve the citing contexts of selected documents. Working in a library, we are in the ideal position to retrieve citing documents and contexts.

Anti 3) Anti-methodological gap

This is central to my idea of individual citation analysis. The anti-methodological gap spans across two levels: a bibliometric context and a historical context. The former should consist of all the relevant bibliometric literature pertaining to the subject(s) area covered by the researcher's curriculum, with a particular attention to the biases and limitations of the indicators included in the analysis in light of the past and current bibliometric debate. Even more importantly, the contextualization will be entrusted with the task of tethering bibliometric indicators to the real world of the scholar's activity through a concise historical reconstruction of the research domains covered by his or her curriculum. The report should cover as many aspects of the history of the discipline as its intellectual foundations and preferential modes of communication, its social and institutional stratification at both the local and international level, past and recent breakthroughs, open questions and puzzle-solving activities, areas of fast-moving development and areas of stagnation.

Of course, in the world of the quick-and-dirty where most bibliometric indicators dwell, the introduction of slow-and-clean procedures might sound as pure utopia. Indeed, there are many difficulties in the realization of a historically informed citation report. The author(s) of the report should have some competence in the scientific field along with basic historiographic skills. Besides, this kind of research is time consuming, you don't just push the button to get a shortcut for quality, and there is a serious risk of relativistic accommodation: the deeper you go in what a scientist does, the more what he or she achieves appears good in its own context. Nevertheless, I believe that the marriage between bibliometrics and context-driven research assessment is worth planning, and that for a micro-level evaluation to be meaningful, the two dimensions of quantity and quality should meet halfway. The library can be a good place for the meeting and maybe also the engagement ceremony. Time will tell if it is true love or just a convenient arrangement.

References

Price, D.J.D., 1986. *Little Science, Big Science...and Beyond*, New York: Columbia University Press.

Pritchard, A., 1969. Statistical bibliography or bibliometrics? *Journal of Documentation*, 25(4): 348-349.