

NOTE DE RECHERCHE

La fièvre de l'évaluation de la recherche. Du mauvais usage de faux indicateurs

Yves Gingras

2008-05

Adresse postale

CIRST
Université du Québec à Montréal
C.P. 8888, Succursale Centre-ville
Montréal (Québec)
Canada, H3C 3P8

Adresse civique

CIRST
Université du Québec à Montréal
Pavillon Thérèse-Casgrain, 3e étage
455, boul. René-Lévesque Est, Bureau W
-3042
Montréal (Québec) Canada
H2L 4Y2

Pour nous joindre

Téléphone : (514) 987-4018
Télécopieur : (514) 987-7726
Courrier électronique : cirst@uqam.ca
Site Internet : www.cirst.uqam.ca

Conception graphique : Marie-Andrée Desgagnés

ISBN 978-2-923333-39-7

Dépôt légal - Bibliothèque et Archives nationales du Québec, 2008

Dépôt légal - Bibliothèque et Archives Canada, 2008

Table des matières

1. Introduction	1
2. Bref historique de la bibliométrie.....	3
3. Quelques propriétés essentielles d'un bon indicateur.....	5
3.1 Adéquation à l'objet.....	6
3.2 Homogénéité de la mesure	7
3.3 Respect de l'inertie propre de l'objet	7
4. Le palmarès de Shanghai	8
5. Le « h index » des chercheurs	10
6. Quelques exemples de bonnes pratiques.....	12
7. Conclusion	14

1. Introduction

Depuis quelques années, le monde académique européen semble atteint d'une véritable fièvre d'évaluation de la recherche et des universités. En France en particulier, la Loi sur l'autonomie des universités et la réforme de l'organisation de la recherche ont créé un climat particulier qui a exacerbé la sensibilité aux questions d'évaluation, peu discutées jusque-là dans le monde académique. Or, l'absence de balises et de réflexions méthodologiques sérieuses donne lieu à ce qu'il faut bien appeler des utilisations anarchiques, pour ne pas dire sauvages, de la bibliométrie, méthode de recherche qui consiste à utiliser les publications scientifiques comme indicateur de la production scientifique. Ajoutées aux habituels (et donc mieux connus) investissements en recherche et développement (R&D) qui sont des mesures d'*input* (intrants) de la recherche, les publications servent de mesure d'*output* (extrants) – et les citations qu'elles reçoivent constituent un indice de leur visibilité internationale et, indirectement, de leur « qualité » et de leur « impact » scientifique. On ne compte plus les lettres publiées dans les revues *Nature* et *Science* ou sur les blogs par des scientifiques qui s'improvisent « évaluateurs » – de leurs concurrents surtout... Ainsi, les scientifiques se lancent leur « h index » au visage pendant que les universités font de même avec leur « position » au classement dit de Shanghai. En France, même le Sénat a cru devoir se mêler de la question pour y voir plus clair et la ministre de l'Enseignement supérieur et de la Recherche, Mme Valérie Pécresse a confirmé que « la réalisation d'un classement européen des universités est une priorité pour la Présidence française de l'Union européenne en matière d'enseignement supérieur »¹.

Comme le montre très bien le rapport du sénateur Joël Bourdin, les différents classements ont d'importantes limites et manifestent la fâcheuse tendance à valoriser systématiquement les universités de certains pays : « le classement de Shanghai est très favorable aux universités américaines... le classement anglais, quant à lui, favorise mieux les performances des établissements du Royaume-Uni... et le classement de Leiden donne de belles places aux

¹ Communiqué de Valérie Pécresse, 2 juillet 2008, www.enseignementsup-recherche.gouv.fr. Voir aussi le Rapport d'information fait au nom de la délégation du Sénat pour la Planification sur le défi des classements dans l'enseignement supérieur. Annexe au procès-verbal de la séance du 2 juillet 2008. Connu sous le nom de rapport Bourdin car signé par le sénateur Joël Bourdin. Copie accessible sur le site : <http://www.senat.fr/rap/r07-442/r07-4421.pdf>

universités néerlandaises... »². Et l'auteur aurait pu ajouter que le classement de l'École des Mines favorise les grandes écoles françaises³...

Cependant, la plupart des critiques se résument à faire ressortir les « limites » des classements sans jamais poser clairement la question préalable de leurs fondements épistémologiques : les indicateurs choisis ont-ils bien la signification qu'on leur attribue? Si ce n'est pas le cas, alors nul besoin de simplement tenter de parer à leurs « limites »; il faut plutôt les remplacer par d'autres, plus adéquats. Car utiliser des classements fondés sur de mauvaises mesures, pourrait avoir des effets pervers en stimulant des politiques fondées sur des problèmes mal identifiés. Ainsi, il est pour le moins curieux d'apprendre que 61% des dirigeants d'établissements d'enseignement supérieur français disent vouloir améliorer leur rang dans le classement de Shanghai alors qu'ils ne savent probablement pas exactement ce qu'il mesure vraiment!⁴ Et on peut s'inquiéter d'entendre Valérie Pécresse déclarer que « Les résultats [pour la France] du classement de Shanghai [,,] plaident pour une politique de regroupement de nos forces »⁵, sans s'assurer que ce classement soit bien valide. Or, comme on le verra plus loin, il ne possède en fait aucune des propriétés que doit posséder un bon indicateur et il serait hasardeux de s'en servir comme guide.

La manie des classements a récemment aussi atteint les revues savantes et l'European Science Foundation (ESF) a publié un classement des revues par discipline, attribuant des cotes A,B, et C selon que les revues sont internationales, nationales ou locales⁶. Or, certains critiquent avec raison un tel classement, mais ils confondent parfois ce type de classement avec de la « bibliométrie », alors qu'il est en fait fondé sur un panel d'experts choisis on ne sait trop comment et qui jugent eux-mêmes de la qualité relative des revues, ce qui est donc subjectif et difficile à contrôler. Les rédacteurs des revues en histoire et sociologie des sciences se sont d'ailleurs concertés pour dénoncer ces classements superficiels et unidimensionnels et plus ou

² Rapport Bourdin, p. 53.

³ Voir www.ensmp.fr/PR/defclassementEMP.pdf

⁴ Rapport Bourdin, p. 98.

⁵ *Les Cahiers de la compétitivité*. Spécial Enseignement supérieur, p. II; encart dans *Le Monde* du 21 mai 2008.

⁶ Voir <http://www.esf.org/research-areas/humanities/research-infrastructures-including-erih/erih-governance-and-panels/erih-expert-panels.html>.

moins occultes quant à leur méthode⁷. Cela indique toutefois la confusion qui existe entre *évaluation* et bibliométrie, confusion qu'il faut lever.

L'intérêt de bien comprendre les propriétés des indicateurs bibliométriques tient au fait qu'il est impossible d'échapper aux évaluations et qu'il est donc plus efficace de critiquer d'abord rigoureusement les indicateurs mal construits, et dont l'utilisation peut engendrer des effets pervers, pour ensuite montrer qu'il est possible de construire des indicateurs utiles et dont la construction est contrôlée.

Il vaut donc la peine de rappeler brièvement les origines de la bibliométrie avant de présenter les propriétés élémentaires que doivent posséder de bons indicateurs (de quoi que ce soit). On pourra alors montrer que les deux indicateurs de la « qualité de la recherche » les plus utilisés actuellement par les décideurs (le classement de Shanghai) et par les chercheurs eux-mêmes (l'index h) n'ont en fait aucune des propriétés nécessaires pour assurer qu'ils indiquent bien ce qu'ils sont censés indiquer.

2. Bref historique de la bibliométrie

Les usages « sauvages » de la bibliométrie qui se multiplient dans la communauté scientifique depuis quelques années ont engendré, avec raison, toute une série de critiques. Toutefois, certaines tendent à « jeter le bébé avec l'eau du bain », en confondant les usages simplistes de cet outil avec l'outil lui-même qui, pourtant, peut être très utile lorsqu'utilisé selon les règles de l'art.

À ses débuts, la scientométrie, qui porte sur la mesure de l'activité scientifique ou la bibliométrie si on se limite aux publications (les deux termes sont devenus pratiquement interchangeables), relevait d'une petite communauté assez méconnue composée de bibliothécaires, sociologues, historiens ou statisticiens, qui étudiaient les transformations du système de la recherche à travers les propriétés des publications scientifiques et des références qu'elles contiennent (les « citations »). Ils faisaient, et font encore, connaître leurs travaux essentiellement par le biais

⁷ Voir www.sauvonsluniversite.com/spip.php?article591.

d'une revue, *Scientometrics*, fondée en 1978. Une trentaine d'années plus tard, les choses ont bien changé et cette « science de la science » est maintenant devenue un enjeu politique important lié à l'évaluation institutionnelle de la recherche et des chercheurs. Depuis le début des années 1990, une revue savante *Research Evaluation* présente des travaux d'évaluation de la recherche fondés sur la bibliométrie. Une caractéristique importante de ces travaux est qu'ils se font essentiellement à une échelle agrégée (grand laboratoire, université, pays) et évitent l'usage de la bibliométrie à des fins d'évaluation des individus. Au fil des décennies, les propriétés agrégées des publications et des citations ont ainsi été établies et sont utiles pour comprendre les transformations de la science et pour remettre en question de nombreux lieux communs auxquels les scientifiques eux-mêmes peuvent souvent adhérer⁸.

Pendant longtemps, seule la compagnie Thomson Reuters (qui a absorbé la compagnie ISI, fondée par Eugene Garfield et qui fut à l'origine de la bibliométrie au début des années 1960⁹) proposait des bases de données bibliographiques (les bases SCI, SSCI et AHCI, pour les sciences, les sciences sociales et les arts et humanités respectivement; elles sont regroupées dans le « Web of science », cumulant aujourd'hui environ 9000 revues et accessible dans les bibliothèques universitaires qui souscrivent aux services de Thomson). Ce long monopole explique que la plupart des travaux de scientométrie reposent sur ces trois bases de données. Depuis 2002, une nouvelle banque de données (SCOPUS) mise en marché par la compagnie Elsevier, couvre davantage de revues (environ 16,000 toutes disciplines confondues) et fait directement compétition à Thomson. L'avantage de ces deux sources au plan de l'évaluation est qu'elles sont contrôlées et que l'on connaît la liste des revues qui y sont recensées. Le désavantage, bien sûr, est qu'elles ne sont pas gratuites...

Google Scholar, et même internet au complet, qui sont accessibles gratuitement, servent de plus en plus comme banques de données pour l'analyse bibliométrique (et, par extension,

⁸ Voir, par exemple, V. Larivière, É. Archambault et Y. Gingras, «Long-term variations in the aging of scientific literature: from exponential growth to steady-state science (1900-2004)», *Journal of the American Society for Information Science and Technology*, vol. 59, no 2, 2008, pp. 288–296.

⁹ E. Garfield, « Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas », *Science*, vol. 122, (3159), July 1955, p.108-11. Pour un bref historique, voir P. Wouters, « Aux origines de la scientométrie. La naissance du Science Citation Index », *Actes de la recherche en sciences sociales*, no 164, 2006, pp. 11-22.

« webométrique »). Le problème avec ces deux sources est qu'elles sont non contrôlées et non reproductibles car leur contenu varie constamment et l'on n'a aucune idée des critères d'inclusion des documents (en fait, il n'y en a pas...), contrairement aux bases du Web of science et de SCOPUS dont la liste des revues recensées est connue, même si elle augmente dans le temps pour tenir compte des nouvelles revues. L'accès gratuit et non contrôlé à Google Scholar et internet, contrairement à l'accès payant à SCOPUS et du Web of Science, a bien sûr contribué à ce climat d'anarchie évaluative, tout chercheur pouvant tenter de « mesurer » sa « visibilité » ou sa « qualité » en concoctant un indicateur spontané fondé sur le nombre de « hit » sur le web ou dans Google Scholar. Par conséquent, il y a récemment eu multiplication des usages « spontanés » de l'évaluation et création de soi-disant indicateurs de l'impact de la recherche qui contribuent à créer un certain chaos dans le monde académique dont les membres ne savent pas toujours comment juger la qualité de ces « mesures ». Ces indices contribuent à la création de palmarès qui fonctionnent comme des « boîtes noires » non analysées, utilisées comme des « faits » solides dont devraient tenir compte les « décideurs » dans leur politique de la recherche.

3. Quelques propriétés essentielles d'un bon indicateur

En fait, l'existence et la persistance de ces indicateurs et palmarès semblent l'effet d'une loi sociale qui veut que, comme on dit en anglais, « any number beats no number » : mieux vaut n'importe quel chiffre que pas de chiffre du tout! Prenons deux exemples de ces indices : le premier fascine les décideurs politiques et les présidents d'université, car il propose un classement mondial des universités (le classement dit « de Shanghai »), réponse miracle à la question de la qualité relative des institutions dans un contexte de « compétition mondiale ». Le second circule plutôt parmi les scientifiques eux-mêmes et « évalue » les chercheurs individuels (le « h index »). Enfin, notons que dans les deux cas, on utilise *un seul nombre* pour classer et « évaluer » la « qualité » de la recherche des individus et des organismes, et ce, malgré le caractère multidimensionnel de la recherche.

Mais pour mieux évaluer, à notre tour, la valeur de ces indicateurs, rappelons quelques propriétés de base d'un indicateur bien construit. Tout d'abord, un « indicateur » est une variable mesurable qui vise à appréhender un concept (ce peut être celui « d'inflation », de « productivité » ou de

« qualité de la recherche »). L'indicateur n'est pas le concept lui-même, mais une façon approchée de mesurer dans le temps les changements du concept visé. Il doit donc être fortement corrélé aux caractéristiques présumées du concept lui-même.

3.1 Adéquation à l'objet

Une première propriété d'un bon indicateur est son adéquation à l'objet. La mesure est-elle appropriée à l'objet évalué? Les résultats que produit l'indicateur sont-ils du bon ordre de grandeur étant donné ce que l'on en sait par ailleurs? Ainsi, le niveau d'investissement en R&D est une bonne première mesure de l'intensité de la recherche dans un pays. Mais supposons que l'on veuille mesurer l'impact scientifique d'un auteur. On peut bien sûr faire un sondage. On peut aussi penser que les citations pourraient servir *d'indicateur* d'un tel impact. Mais il ne suffit pas de le décréter de façon tautologique; il faut d'abord tester ce lien en trouvant une relation entre une mesure *indépendante* et la mesure donnée par les citations. Or, les travaux de sociologie des sciences et de bibliométrie depuis les années 1970 ont maintes fois montré qu'une telle corrélation existe entre le niveau de citation et la renommée mesurée par des prix obtenus ou des nominations académiques¹⁰. En fait, le mythe des grands savants peu cités est bien cela : un mythe. Einstein, par exemple, est très rapidement hautement cité dans les 4 ans suivant ses publications de 1905. Le but n'est pas alors de savoir si Einstein a 100 ou 104 citations, mais de voir si son niveau de citation est effectivement de loin supérieur à la moyenne, ce qui est d'ailleurs le cas¹¹. Notons cependant que cet indice des citations a été surtout validé dans des domaines des sciences de la nature et donc on ne peut le transférer *sans précautions* dans les secteurs des sciences sociales et encore moins dans les lettres et les sciences humaines, car ces dernières disciplines utilisent davantage le livre que l'article comme mode de diffusion de leurs résultats¹². En somme, s'assurer que l'indicateur est bien adéquat à l'objet exige d'effectuer des tests et d'analyser les modes de production des savoirs en vigueur dans les différentes disciplines.

¹⁰ J.R. Cole et S. Cole, *Social Stratification in Science*, Chicago, Chicago University Press, 1973.

¹¹ Sur le cas Einstein, voir Y. Gingras, « The Collective Construction of Scientific Memory: The Einstein-Poincaré Connection and its Discontents, 1905-2005 », *History of Science*, vol. 46, no 151, March 2008, pp. 75-114.

¹² Sur cette question voir Vincent Larivière, É. Archambault, Y. Gingras, É. Vignola-Gagné, « The place of serials in referencing practices: Comparing natural sciences and engineering with social sciences and humanities », *Journal of the American Society for Information Science and Technology*, vol. 57, no 8, June 2006, pp. 997-1004; É. Archambault, É. Vignola-Gagné, G. Côté, V. Larivière, Y. Gingras, « Benchmarking Scientific Output in the

3.2 Homogénéité de la mesure

Une seconde propriété importante qu'un bon indicateur doit posséder est d'être homogène dans sa composition. Dans le domaine de la recherche par exemple, un indicateur homogène (à l'échelle d'un pays par exemple) de l'intensité de l'activité de recherche est fourni par le nombre d'articles publiés dans les principales revues scientifiques. Il s'agit là d'une mesure d'*output* (extrants) qui peut aussi être comparée à une mesure d'*input* (intrants), la valeur des investissements en recherche (mesurée en euro). Ces indicateurs permettent de comparer les pays et même les institutions entre elles. Ils peuvent aussi servir à construire une *cartographie* descriptive des activités selon deux mesures différentes : les intrants et les extrants. On peut aussi construire le rapport de ces deux mesures et obtenir un indice composite de productivité (*input/output*). Par contre, si l'on prend cet indicateur comme une mesure de « qualité » de la recherche (et non de son efficacité) ou qu'on le combine avec un indicateur de réputation fondé sur un panel d'experts, alors on obtient un indicateur assez hétérogène et qui pourra varier de façon imprévue, sans pour autant avoir une signification claire.

3.3 Respect de l'inertie propre de l'objet

Une dernière propriété importante d'un indicateur est qu'il doit varier en conformité avec l'inertie propre de l'objet mesuré, certaines choses variant plus ou moins rapidement, selon leurs caractéristiques intrinsèques. Prenons l'exemple du thermomètre et supposons qu'au lieu de la bonne vieille colonne de mercure ou d'alcool on utilise un instrument électronique à écran numérique et que l'on observe dans une pièce sans courant d'air qu'il indique 20 degrés, puis une minute plus tard 12 et encore une minute plus tard, 30. Il est certain que le bon sens forcera l'observateur à conclure que l'instrument est défectueux, car l'on sait très bien (et la thermodynamique le confirme) que la température de la pièce ne peut varier aussi rapidement en trois minutes! Or, il est bien connu que les grandes institutions académiques sont de lourds paquebots qui ne changent pas de cap très rapidement (et c'est très bien ainsi, car cela leur permet d'éviter de « répondre » à des demandes éphémères, voire frivoles¹³). En conséquence, un

Social Sciences and Humanities: The Limits of Existing Databases, *Scientometrics*, vol. 68, no 3, 2006, pp. 329-342.

¹³ À cet égard, il est absurde d'affirmer que « la formation doit rester connectée avec un marché du travail extrêmement volatil » car la formation prend des années, ce qui est évidemment incompatible avec la « volatilité »

palmarès annuel qui montrerait qu'une institution est passée en une seule année du 1^{er} au 6^e rang ou du 8^e au 2^e rang suggérerait fortement que l'indicateur utilisé est défectueux ou trop imprécis, et non pas que la qualité de l'institution a chuté! De plus, étant donné la variance naturelle des données d'une année à l'autre, il est clair que la plupart des changements annuels de rang observés dans les palmarès sont en fait aléatoires et n'ont aucune signification réelle. Aux États-Unis, par exemple, le *National Research Council* produit un classement de tous les programmes de doctorat des universités américaines dans toutes les disciplines. Il le fait une fois tous les dix ans. Pourquoi cette faible fréquence? Parce que, en plus des coûts élevés d'une telle opération, la probabilité qu'un programme académique soit excellent en 2008 et médiocre en 2009 est pratiquement nulle. Cette fréquence respecte donc le fait que l'institution universitaire est passablement inertielle. Cela suggère aussi qu'évaluer de grands groupes tous les 4 ans n'a pas vraiment de sens (et constitue un gaspillage de ressources) et qu'une fréquence de 6-8 ans serait plus réaliste pour observer de véritables changements.

À la lumière de ces trois critères de validité, qui suffisent pour détecter les mauvais indicateurs, jetons maintenant un coup d'œil plus attentif sur le palmarès de Shanghai et sur l'indice h (« h index »). Le choix de s'arrêter à ces deux indices s'impose, car ils offrent deux exemples de classements construits à des échelles différentes : celui de Shanghai évalue des institutions tandis que l'indice h évalue des individus. On le devine déjà : ces deux mesures ne constituent nullement des indicateurs valables et ils devraient être abandonnés au plus vite par ceux qui se targuent de prendre des décisions à la lumière d'indicateurs « objectifs » et « internationaux »...

4. Le palmarès de Shanghai

Ce classement des supposées « meilleures » universités mondiales est actuellement très en vogue et semble attendu chaque année avec impatience par de nombreux dirigeants d'universités. Il est composé de la somme de six mesures dont quatre ont un poids de 20% chacun ; 1) membres du corps universitaire ayant reçu un Nobel ou une médaille Fields (pour les mathématiques); 2) nombre de chercheurs de l'institution parmi la liste des « plus cités » de Thomson Reuters; 3)

du marché! D'où l'importance d'une formation de base qui transcende cette « volatilité »... Voir *Les cahiers de la compétitivité*, p. 1, cité en note 2.

nombre d'articles de l'institution publiés dans les revues *Nature* et *Science*; 4) nombre total d'articles recensés dans le Web of Science de la compagnie Thomson Reuters. Deux autres mesures ont un poids de 10% chacun : 5) nombre d'anciens étudiants ayant reçu un Nobel ou une médaille Fields; 6) ajustement des résultats précédents selon la taille de l'institution. Il est tout à fait évident que l'indice final de ce palmarès se fonde sur la somme de plusieurs mesures hétérogènes, car le nombre de publications dans *Science* et *Nature* n'est pas commensurable au nombre de prix Nobel. Chose plus surprenante, il a été montré que les données sur lesquelles il est fondé sont difficilement reproductibles¹⁴. On pourrait aussi mettre en cause le choix d'un indicateur comme le nombre d'articles dans *Science* et *Nature* quand on sait que ces deux revues ont un très fort biais américain : 72% des articles parus dans *Science* en 2004 ont au moins une adresse américaine, et cette proportion est de 67% dans la revue britannique *Nature*. Surtout, au regard du critère de l'inertie, il y a lieu de se poser de sérieuses questions sur la validité d'un indice qui fait varier la position d'une université de plus de 100 rangs dans le palmarès par le seul fait d'attribuer à l'université de Berlin ou à l'université Humboldt le Prix Nobel d'Einstein obtenu en 1922! Sans parler du fait que l'on peut se demander si la qualité d'une université en 2006 peut être influencée de la sorte par des travaux effectués plus de 80 ans auparavant¹⁵. On a noté plus haut que certains voient dans ce classement la preuve qu'il faut construire de gros regroupements pour être plus « compétitifs ». Or, c'est oublier que le California Institute of Technology (Caltech) se classe au 6^e rang dans le classement de Shanghai alors qu'il est une institution de très petite taille (environ 300 professeurs et un peu plus de 2000 étudiants). Et l'université de Princeton qui elle aussi de taille relativement réduite, avec environ 5,000 étudiants, se classe en 8^e position. Vu sous cet angle, cela ne conforte sûrement pas l'idée des grands regroupements et pourrait plutôt inciter à conclure que « small is beautiful »...

En fait, seule une psychosociologie des dirigeants universitaires et autres fonctionnaires ministériels haut placés pourrait expliquer un tel engouement pour un classement qui n'a en réalité aucune valeur scientifique. Il est aussi probable que l'importance soudaine accordée à ce classement, soit un effet des discours sur l'internationalisation du « marché universitaire » et de

¹⁴ R.V. Florian, « Irreproducibility of the results of the Shanghai academic ranking of world universities » *Scientometrics*, vol. 72, July 2007, pp. 25-32.

¹⁵ Sur la controverse tragi-comique entre ces deux universités et les responsables du classement de Shanghai voir *Science*, 24 August 2007, p. 1026.

la recherche de clientèles étrangères lucratives qui viendraient ainsi combler les revenus insuffisants provenant des gouvernements. De nombreux dirigeants universitaires qui envoient des délégations en Chine semblent y voir en effet un « marché » potentiellement lucratif qu'il ne faudrait pas laisser aux seules universités américaines. Enfin, il sert aussi de façon stratégique les acteurs qui veulent réformer le système universitaire et se servent de ces classements de façon opportuniste pour justifier leurs politiques. En fait, il est même probable que dans l'éventualité où les universités françaises se seraient très bien classées, il aurait été plus difficile de justifier les réformes actuelles et que les décideurs auraient alors jeté un regard plus critique sur un tel classement qui, à l'inverse, aurait été bien reçu par ceux qui s'opposent aux réformes.

5. Le « h index » des chercheurs

Analysons maintenant un autre indicateur à la mode, mais qui circule plutôt parmi les chercheurs que dans les bureaux de présidents d'universités. L'indice h a été construit par le physicien Jorge E. Hirsch, de l'université de Californie à San Diego¹⁶. L'indice h d'un chercheur est défini comme étant égal au nombre d'articles n qu'il a publié et qui ont reçu au moins n citations (pour une période donnée). Par exemple, un auteur qui a publié 20 articles parmi lesquels 10 ont au moins 10 citations chacune aura un indice h de 10. Cet indicateur de « qualité » de la recherche d'un individu est donc un composite de la production (nombre d'articles écrits) et de la « visibilité » (nombre de citations reçues) et non pas, comme le dit son auteur une mesure homogène d'*output*, c'est-à-dire d'extrait. Un tel mélange devrait déjà nous faire douter de la fiabilité d'un tel indice. Mais, comme s'il contribuait à satisfaire le narcissisme des scientifiques, son usage s'est généralisé parmi eux en moins de deux ans, et est même incorporé dans certaines banques de données! Selon son concepteur, cet indice favoriserait « une évaluation plus démocratique de la recherche »¹⁷... Or, cet index n'a que peu d'utilité au niveau individuel et son usage au plan de la prise de décision est inutile, car en fait il est très fortement corrélé au nombre

¹⁶ J.E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National Academy of Sciences*, 102(46), 2005; pp. 16569-16572. Malgré son titre, l'indice ne mesure pas vraiment « l'output » mais un mélange d'*output* et de « visibilité ».

¹⁷ Voir <http://pubs.acs.org/cen/science/86/8621sci1.html>.

total d'articles et est ainsi redondant¹⁸ en plus de pouvoir générer des effets pervers. Un exemple simple suffit à le démontrer.

Comparons deux cas de figure : un jeune chercheur a publié seulement trois articles, mais ceux-ci ont été cités 60 fois chacun (pour une période de temps donnée); un second chercheur, du même âge, est plus prolifique et possède à son actif 10 articles, cités 11 fois chacun. Ce second chercheur a donc un indice h de 10, alors que le premier a un indice h de 3 seulement. Peut-on en conclure que le second est trois fois « meilleur » que le premier? Bien sûr que non....

En somme, la bibliométrie appliquée au niveau individuel peut être dangereuse; aussi les experts en bibliométrie en évitent-ils les aléas en limitant généralement son usage à l'analyse du comportement d'agrégats (université, région, pays). Ces résultats peuvent avoir une valeur statistique, tandis que les données individuelles varient considérablement et ce d'autant plus que les lois de distribution des publications et des citations ne sont pas des lois normales, mais plutôt log-normales ou encore des lois de puissance de type Pareto, ce qui fait que la variance est très souvent plus élevée que la moyenne – Alfred J. Lotka a montré dès 1926 que la production scientifique suivait une courbe de type Pareto, c'est-à-dire qu'une minorité de chercheurs sont très productifs alors que la majorité des chercheurs produit peu¹⁹.

Il vaut la peine de rappeler, contre ceux qui croient que ce ne sont que les dirigeants des institutions qui veulent imposer ce genre d'indicateurs que, dans le cas de l'indice h , sa propagation rapide dans certaines disciplines scientifiques s'est faite à la base. Ce sont en effet les scientifiques eux-mêmes qui succombent souvent aux usages anarchiques de la bibliométrie individuelle et qui, siégeant parfois sur différents comités et conseils d'administration d'organes décisionnels de la recherche, suggèrent d'en généraliser l'usage. Cela confirme que dans le champ scientifique, « l'ennemi » est souvent moins le « bureaucrate » que le collègue...

¹⁸ Van Leeuwen, Th. N. "Testing the validity of the Hirsch-index for research assessment purposes", *Research Evaluation*, June 2008, vol. 17, p. 157-160.

¹⁹ Alfred. J. Lotka, "The frequency distribution of scientific productivity", *Journal of the Washington Academy of Science*, vol. 36, 1926, p. 317-323.

6. Quelques exemples de bonnes pratiques.

Cela étant dit, il est possible de construire des indicateurs agrégés de la recherche qui donnent une bonne idée de la position relative des universités et des pays dans un système de recherche, national ou mondial. Le plus ancien est sans doute la proportion du PIB d'un pays consacré à la recherche et au développement (R&D), publié depuis longtemps par l'OCDE et qui permet une certaine mesure comparée des niveaux d'activités de R&D des pays. De même, le nombre total d'articles publiés dans les principales revues scientifiques recensées dans les bases de données du Web of Science ou de SCOPUS donne aussi un classement utile. En fait, malgré les différences de couverture bibliographique, ces deux bases de données produisent essentiellement le même classement, au moins pour les 25 pays les plus importants²⁰. De plus, il existe une corrélation très forte entre le nombre d'articles produit par pays et son niveau de dépenses en R&D. En termes de nombre total de publications produites en 2005 par les huit pays les plus productifs²¹, par exemple, la France se classe au 6^e rang. On peut faire un pas de plus et calculer un indice de la visibilité de ses articles, mesurée par le nombre de citations reçues par ces articles sur une période de deux ans suivant leur publication (on pourrait le faire pour 3 ou 5 ans au besoin) en normalisant pour tenir compte des différents taux de citations par champ de recherche. Selon cette mesure, qui notons-le est homogène et distincte du nombre de publications, la France monte alors au 5^e rang, toujours pour l'année 2005. La différence s'explique aisément : pour le nombre de publications, la Chine passe avant la France, mais pas (encore) pour ce qui concerne les citations reçues par ses articles. De même, le Canada se classe au 8^e rang pour la production, mais passe au 4^e rang en termes de citations.

Étant donné la diversité des champs de recherche, le mieux est encore de construire ces indices par grands domaines, car toutes les universités et tous les pays ne sont pas également actifs, ni visibles, dans tous les secteurs. Ainsi, dans le secteur de la biologie, la France est en 4^e position en termes de citations relatives et, en mathématiques, sa position oscille, entre 2000 et 2005, entre

²⁰ E. Archambault, D. Campbell, Y. Gingras, V. Larivière, "WOS vs. Scopus: On the reliability of scientometrics", à paraître dans le *Book of Abstracts of the 10th International Conference on Science and Technology Indicators*, 2008.

²¹ On pourrait prendre l'ensemble des pays ce qui changerait bien sûr l'ordre donné ici à titre d'exemple, car plusieurs petits pays produisent beaucoup moins d'articles (la Suisse par exemple), mais se classent très haut en termes de citations.

la 2^e et la 3^e position, ce qui confirme la forte tradition des mathématiques en France²². On pourrait ainsi continuer à décliner les différents champs, mais le but était simplement ici de suggérer que construire des indices simples et homogènes permet de les interpréter plus facilement et de voir s'ils changent de façon continue ou abrupte²³. Ainsi, la fluctuation autour d'un seul rang d'une année à l'autre est en fait due à la variance naturelle des données et ne doit pas donner lieu à une surenchère d'interprétation. Par contre, un déclin régulier sur plusieurs années consécutives (disons 5 ans) d'un indicateur (soit en termes de nombre d'articles ou de citations relatives) devrait donner lieu à un travail d'interprétation plus sérieux. Ainsi, la forte montée de la Chine en termes de production globale – elle passe de la 8^e position en 2000 à la 4^e en 2005 – signifie bien une stratégie de production accrue dans les revues internationales, mais elle ne s'est pas encore traduite par une remontée importante en termes de citations.

La National Science Foundation publie d'ailleurs le *Science and Engineering Indicators* depuis 1972, lequel comprend des données bibliométriques comparatives très utiles. De même, l'Institut de la statistique du Québec, en collaboration avec l'Observatoire des sciences et des technologies (OST) de l'UQAM, publie de façon régulière des données bibliométriques comparées sur le Québec, le Canada et quelques autres pays²⁴. En France même, l'Observatoire des sciences et des techniques (OST), créé en 1990, fait lui aussi paraître chaque deux ans un volume intitulé *Indicateurs des sciences et des technologies* qui inclue des données bibliométriques²⁵. Ces exemples montrent qu'il est possible d'éviter les effets les plus simplistes des classements en organisant adéquatement les données. Bien sûr, il est probablement impossible d'éviter complètement l'effet de classement, car quelqu'un quelque part peut toujours jouer au jeu de l'évaluation sauvage. Mais quand les institutions auront appris à se servir d'indicateurs bien construits et à les interpréter dans leurs limites de validité, les dangers de dérapage seront moins grands en termes de prise de décision. Car la production et la publication de tableaux

²² Ces données ont été compilées par l'OST-UQAM.

²³ Notons tout de même, qu'une analyse par grands champs de recherche entre 2000 et 2005, montre que sur l'indice des citations relatives, la France reste toujours en très bonne position *parmi les 8 pays qui produisent le plus grand nombre d'articles*, car elle oscille entre les positions 5 et 6 en physique, sciences biomédicales, médecine clinique et sciences de la Terre et de l'espace. Elle occupe la seconde ou troisième position en sciences de l'ingénieur et en mathématiques, toujours en termes de citations relatives sur deux ans. Rien dans tout cela ne suggère une crise de la recherche française que ceux qui ont les yeux rivés sur le palmarès de Shanghai semblent détecter...

²⁴ Voir par exemple, Institut de la statistique du Québec, *Compendium d'indicateurs de l'activité scientifique et technologique au Québec*, édition 2008, Québec, chapitre 3. Version disponible sur Internet : www.stat.gouv.qc.ca.

²⁵ *Indicateurs des sciences et des technologies* sous la direction de G. Filiatreau, Paris, Economica, 2006.

bibliométriques comparatifs par domaine de recherche, et même par institution, compilées sur quelques années pour renforcer la valeur statistique des données, peuvent permettre d'analyser sereinement les tendances mondiales ou nationales de la recherche et prendre des décisions éclairées en fonction des priorités locales ou nationales.

7. Conclusion

Que faire donc pour contrer cette anarchie bibliométrique qui gagne du terrain? Il faut d'abord s'appliquer à convaincre les institutions de cesser d'utiliser des indices fondés sur des mesures douteuses et construits par des chercheurs qui s'improvisent bibliométriciens le temps de créer un indicateur sur le coin d'une table qui a l'avantage de favoriser ses collègues et de défavoriser ses compétiteurs (cela s'est vu...). Surtout, il est nécessaire d'éduquer et de convaincre les scientifiques eux-mêmes des dangers des mauvais usages de la bibliométrie. Au lieu de laisser des amateurs se convertir en évaluateurs de la recherche sous prétexte qu'ils sont des acteurs de la recherche, mieux vaut laisser l'évaluation de la recherche aux personnes qui y consacrent assez de temps pour pouvoir distinguer les pseudo-métriques des indicateurs robustes. On diminuera probablement ainsi les chances de dérapage et les effets pervers générés par des mesures et des classements de la recherche qui peuvent à court terme servir des intérêts stratégiques, mais qui à long terme ne pourront qu'impulser des réformes qui aboutiront à rien, ou pis encore, seront la source de nouveaux problèmes, ayant été conçues à partir des mesures qui ne correspondaient pas à la réalité. Et quant aux évaluations des chercheurs individuels, les comités de pairs composés d'experts sont encore les mieux placés pour juger de la quantité et de la qualité du travail accompli. Rien n'interdit bien sûr de consulter des indices bien construits, mais le plus souvent ils ne font que confirmer ce que des chercheurs bien au fait des activités de leurs collègues savaient déjà. Au niveau agrégé, par contre, les indicateurs bibliométriques donnent des informations qu'aucun acteur – au regard toujours fixé à son pare-brise quand ce n'est pas à son rétroviseur – n'est en mesure de vraiment connaître.

En somme, en s'assurant de confier le travail à de véritables experts en évaluation quantitative de la recherche qui savent construire et surtout interpréter correctement les données à l'intérieur de leur domaine de validité, il est possible de faire des évaluations qui peuvent être vraiment utiles

comme aide à la prise de décision. Bien sûr, tout comme on peut toujours débattre de la signification précise de la variation du taux de chômage ou de l'inflation, on pourra diverger sur les conclusions à tirer de tels indicateurs bibliométriques.

Cependant, l'on aura au moins construit des indicateurs pertinents et guéri ainsi la fièvre évaluative en éradiquant le redoutable virus qui l'a causée : l'improvisation.

Titres parus

- 08-04 **Beaudry**, Catherine et Ruby **Farcy**, Dynamiques d'innovation et politiques de financement en biotechnologie.
- 08-03 **Hanel**, Petr, Productivity and Innovation: An Overview of the Issues
- 08-02 **Hanel**, Petr, Skills Required for Innovation : A Review of the Literature
- 08-01 **Monchatre**, Sylvie, «L'«approche par compétence», technologie de rationalisation pédagogique. Le cas de la formation professionnelle au Québec
- 07-07 **Gentzoglani**s, Anastasios, «Technological and Regulatory Changes in the Financial Industry in the MENA Region: Competitiveness and Growth
- 07-06 **Larivière**, Vincent, Alesia **Zuccala** et Éric **Archambault**, «The Declining Scientific Impact of Theses : Implications for Electronic Thesis and Dissertation Repositories and Graduate Studies»
- 07-05 **Doray**, Pierre, Lucia **Mason** et Paul **Bélanger**, «L'art de vaincre l'adversité : le retour aux études des adultes dans l'enseignement technique»
- 07-04 **Chenard**, Pierre, Éric **Francoeur** et Pierre **Doray**, «Les transitions scolaires dans l'enseignement postsecondaire : formes et impacts sur les carrières étudiantes»
- 07-03 **Proulx**, Serge, Julien **Rueff** et Nicolas **Lecomte**, «Une appropriation communautaire des technologies numériques de l'information»
- 07-02 **Gentzoglani**s, Anastassios, «International Competitiveness in the Telecommunications and ICT Sectors : A Cross Country comparison»
- 07-01 **Gentzoglani**s, Anastassios, «Innovation, réglementation et choix organisationnels au niveau du marché électrique : le cas des pays en Afrique»
- 06-06 **Prud'homme**, Julien, «Histoire de l'école d'orthophonie et d'audiologie de l'Université de Montréal, 1956-2002. Des luttes professionnelles à l'épanouissement disciplinaire»
- 06-05 **Banik**, Marc, «Regulating the Risks of Dietary Supplements: An Economic analysis of Qualified Health Claims and Efficacy Statement Disclaimers»
- 06-04 **Leonard**, Robert, «From Chess to Catastrophe: Psychology, Politics and the Genesis of von Neumann's Game Theory»
- 06-03 **Leonard**, Robert, «From Austro-liberalism to Anschluss: Oskar Morgenstern and the Viennese Economists in the 1930's»
- 06-02 **Banik**, Marc, «Reappraising the Efficiency of Probabilistic Patents and Prescriptions for Patent Policy Reform»
- 06-01 **Proulx**, Serge, Nicolas **Lecomte** et Julien **Rueff** «Portrait d'une organisation québécoise orientée vers l'appropriation sociale des technologies de l'information et de la communication en milieu communautaire»
- 05-06 **Castonguay**, Stéphane, «La dynamique du changement scientifique en contexte d'application : la fondamentalisation de l'entomologie économique aux États-Unis au 20^e siècle»
- 05-05 **Larivière**, Vincent, Éric **Archambault**, Yves **Gingras** et Étienne **Vignola-Gagné**, «The Place of Serials in Referencing Practices: Comparing Natural Sciences and Engineering with Social Sciences and Humanities»

www.cirst.uqam.ca

CIRST

Centre interuniversitaire de recherche
sur la science et la technologie

Le CIRST est, au Canada, le principal regroupement de chercheurs dont les travaux sont consacrés à l'étude des multiples dimensions de l'activité scientifique et technologique. La production régulière de travaux de recherche ainsi que la formation de nouveaux chercheurs contribuent à éclairer les débats et à informer les décideurs sur les enjeux actuels des sciences et des technologies. Ces recherches s'ordonnent autour de trois grands axes : l'analyse du développement scientifique et technologique, l'analyse socioéconomique et la gestion des technologies et enfin, l'analyse sociopolitique des usages et des incidences des technologies.



Créé en 1986, le CIRST est reconnu par trois universités : l'Université du Québec à Montréal, l'Université de Montréal et l'Université de Sherbrooke. Il rassemble une quarantaine de chercheurs en provenance d'une douzaine d'institutions et des disciplines suivantes : histoire, sociologie, science politique, philosophie, sciences économiques, sciences administratives et communications. Le CIRST fournit un milieu de formation par la recherche à de nombreux étudiants aux cycles supérieurs dans les domaines de recherche de ses membres.

