

Laurence COUTROT

laurence.coutrot@ens.fr

Centre Maurice Halbwachs, (CNRS, ENS, Ehess)
Paris, France.

“The use of citation statistics to assess research production in the social sciences.”

The attempt to assess the quality of research using quantitative methods based on citation statistics is a growing trend. The paper presents the methods, identifies the major factors that can account for this development of quantitative tools. The evaluation of institutions, journals and individual researchers is considered. Major flaws appear, not only because of some technical imperfection in the data banks. Citation practices are not the neutral, objective lens one could expect to observe this highly complex social activity of science production. A test conducted in sociology reveals a discrepancy in the data between American and French social scientists. Among the latter, the number of citations is so tiny that any attempt to use quantitative data to assess French sociology production appears highly questionable. Finally, it is shown that citations statistics as a way to assess research products is also the target of criticism on behalf of other disciplines, including branches of hard sciences, and that the same protest against these methods develops all over the planet. Citation statistics cannot serve as a proxy for peer review.

L'EVALUATION EN SCIENCES SOCIALES : VOUS AVEZ DIT « QUANTIFIÉ » ?

Article sous presse (version en italien) à Quaderni di Sociologia,

Laurence COUTROT
laurence.coutrot@ens.fr

CNRS, Centre Maurice Halbwachs, Paris.
Février 2009.

Introduction

J'avance ici sur un terrain miné. Le sujet que j'aborde, celui de la bibliométrie comme outil d'évaluation de la recherche, a été évoqué à répétition depuis les années 1960. Je voudrais suggérer que la question des indicateurs bibliométriques est loin de se limiter à un problème technique. C'est un outil dans une lutte de pouvoir entre le corps politique et celui des chercheurs. Cette guerre n'est pas neuve. Elle a connu des phases tragiques au moment de la seconde guerre mondiale. On a observé un regain il y a une vingtaine d'années.

En 1989 eut lieu, en Italie, sous l'égide du comité pour les affaires scientifiques de l'OTAN, un vaste colloque sur la politique de la recherche réunissant une centaine d'éminents représentants de 23 pays. A cette occasion, Bernard Barber, (Columbia U.), rendant compte du livre collectif qui en est issu, rappelait, à moitié en plaisanterie, mais à moitié seulement, que pour certains « la science est une affaire trop sérieuse pour être laissée aux mains des savants »¹.

Nous sommes dans une nouvelle phase et l'épisode de « quantophrénie » que traverse actuellement la France n'en est que l'un des symptômes : l'espoir de quantifier la production scientifique et d'en évaluer « objectivement » la qualité grâce à des outils bibliométriques est lui aussi récurrent.

Cette attirance pour des indicateurs quantitatifs afin de mesurer la production scientifique est particulièrement forte dans le contexte français actuel. On peut être tenté de rapprocher cette tendance d'autres réformes qui touchent les politiques publiques dans leur ensemble. Ainsi, la LOLF ou loi d'orientation de la législation financière consiste à rendre toute ligne budgétaire imputable à une activité particulière. L'esprit général est de développer toutes les mesures qui permettent de rendre compte de chaque dépense publique particulière. Cette disposition a des conséquences graves dans bon nombre de domaines: santé, affaires sociales, recherche.

A une conception dans laquelle la recherche est un bien public que l'on finance de façon globale, permettant ainsi de développer des recherches fondamentales et d'accumuler du savoir par provision, se substitue l'idée que la recherche est une dépense publique dont il convient de contrôler minutieusement chaque élément. Un autre élément du contexte français actuel est l'idée que la recherche publique ainsi que l'enseignement supérieur doivent être soumis à des évaluations budgétaires strictes. L'espoir étant que des financements d'origine privée viennent prendre la relève de l'argent public. La méfiance de la majorité actuelle et du Président vis-à-vis du monde de la recherche et des universitaires, accusés de s'auto-évaluer, a éclaté dans le discours présidentiel du 22 janvier 2009.

Dans ce contexte, on a vu se développer, au Ministère de la recherche et de l'enseignement supérieur, comme au sein des organismes de recherche eux-mêmes, des cellules de management dans lesquelles des fonctionnaires produisent des indicateurs scientifiques pour

1 (Cozzens (et al.) ed., 1990).

évaluer les institutions et les personnes dans des domaines où ils ne sont pas personnellement compétents. Ils sont alors obligés de forger ou d'utiliser des outils technocratiques «aveugles» dont ils ne saisissent pas nécessairement les limites.

Je commencerai par quelques remarques générales concernant la bibliométrie : je rappellerai les grands principes qui fondent cette démarche, j'évoquerai quelques éléments du contexte de la recherche française qui accompagnent cette fièvre bibliométrique, je décrirai les principales bases de données.

Dans une seconde partie, je distinguerai l'évaluation de trois types d'objets sociaux: les institutions, les revues, les personnes. Dans une troisième partie, je montrerai que cette tentation de quantifier l'évaluation de la recherche n'est pas un problème local ; j'élargirai le champ d'investigation à d'autres disciplines que la sociologie et à d'autres pays que la France.

I. Quelques remarques générales sur les outils bibliométriques.

1. *Les principes : notoriété et qualité*

Considérant que la publication est une trace essentielle de l'activité scientifique et que la « notoriété », le fait d'être cité, est un indicateur de « l'influence » d'un auteur, on observe qui cite qui, qui est cité par qui et combien de fois... On peut ainsi calculer un « facteur d'impact » d'un auteur ou d'un groupe d'auteurs, d'un laboratoire, d'une institution, etc.

Tant qu'il s'agit de repérer les vedettes, cette méthode ne pose aucun problème : la plupart du temps, un chercheur ayant reçu le prix Nobel a beaucoup publié et ses travaux sont fréquemment cités. Malheureusement, ce sont là des cas rares où la mesure bibliométrique est juste mais inutile : on sait que la corrélation entre célébrité et le fait d'avoir reçu le prix Nobel, (ou même d'être susceptible de le recevoir) est bonne.

Le problème est plus délicat lorsqu'on parle non plus de célébrité mais de qualité et que l'on entend utiliser cette méthode pour évaluer la qualité d'une production scientifique, qu'il s'agisse d'institutions, de revues ou de chercheurs².

2. *Le contexte :*

Le jugement traditionnel de la production scientifique repose sur l'évaluation par les pairs : des collègues de rang égal ou supérieur, qui ont lu et sont à même de comprendre au fond la substance et les méthodes d'une publication scientifique. Ce mode d'évaluation semble soudain remis en cause, au moins dans certains cercles. On le juge désormais trop coûteux en temps et certaines autorités suspectent l'impartialité des « pairs » dont le jugement est sollicité.

A cette mise en cause du mode traditionnel d'évaluation, on peut associer plusieurs événements récents qui ont touché la recherche française.

Une nouvelle agence d'évaluation (AERES), composée exclusivement de membres nommés, vient se superposer aux structures traditionnelles d'évaluation (Comité national du CNRS et CNU pour les universités), dans lesquelles le rôle des membres élus par la communauté scientifique est important. Les responsables de cette nouvelle instance sont confrontés à la nécessité d'évaluer les performances de recherche pour un nombre croissant d'individus : outre les activités des chercheurs des EPST (CNRS, INED, etc.), l'Agence tente

²Je reviens plus en détail sur ce point dans la fin de la deuxième partie de l'article.

d'évaluer également celles des maîtres de conférences et professeurs de l'enseignement supérieur, rebaptisés « enseignants-chercheurs ». On assiste d'ailleurs à la création par le Ministère de la recherche d'un terme qui peut laisser rêveur, celui de « chercheur publiant ».

Deuxièmement, la publication du classement de Shangaï qui place les universités françaises à un rang assez médiocre, suscite dans le pays, notamment parmi ses décideurs, une inquiétude probablement démesurée. Des efforts de production d'un modèle d'évaluation alternatifs ont été fournis (Ecole des Mines). A tort ou à raison, cet ensemble de préoccupations a généré un regain d'intérêt pour les modes quantitatifs d'évaluation et l'élaboration d'indicateurs scientifiques.

3. Les sources

Revenons un peu en arrière. La bibliométrie a entamé ses débuts quand un Américain de génie, Eugene Garfield, a entrepris de développer une base documentaire permettant d'étudier les « structures sociales, culturelles, et cognitive latentes de la pratique scientifique »³.

C'est bien à des fins de documentation scientifique, d'histoire et de sociologie des sciences qu'était destiné l'outil initial, qui deviendra plus tard le « Web of Science » (WoS).

Le WoS demeure l'un des principaux outils disponibles pour analyser les publications scientifiques. Il fut initialement développé au début des années 1960 par E. Garfield à l'Institute for Scientific Information (ISI)⁴, entreprise aujourd'hui rachetée par Thomson Reuters, qui a également développé le logiciel « Endnote ». L'entreprise, basée à New York et Philadelphie, recense environ 10.000 revues scientifiques (dont 1.865 pour le Social Science Citation Index) dans 42 langues différentes. Elle emploie plus de 50.000 salariés dans le monde entier.

Les revues recensées sont en principe sélectionnées en fonction de critères de « sérieux ».

Chaque article recensé fait l'objet d'une indexation (par mots-clé), les noms des auteurs sont répertoriés par leur institution de rattachement⁵.

Plusieurs entrées sont disponibles : par périodique ou par nom d'auteur. Pour chaque entrée on dispose de la liste et du nombre d'items repérés (articles, compte-rendu, notes, etc.) du nombre de citations de ceux-ci, des dates, et l'on peut repérer les auto-citations. A partir de ces informations, on peut calculer automatiquement soit le « facteur d'impact » pour une revue ou un ensemble de revues, soit un indicateur dit « h-index » pour les individus.

L'un des modules du WoS, le Journal Citation Research (JCR) fournit un certain nombre d'informations sur les revues regroupées en catégories : nom du périodique, nombre d'articles publiés pour une année donnée, nombre de fois où un article a été cité (par année ou cumulé), liste des périodiques qui ont cité un article de ce périodique pour une année de référence (*Cited search*). On peut ainsi retrouver des articles cités dans des revues non incorporées dans la base de données du WoS. Malheureusement, les bugs sont nombreux : on trouve ainsi dans

3 « To understand how interacting cognitive and social structure of science affect the thought and behavior of scientists » (Merton, 1979)

4 Pour une histoire détaillée de la création de l'ISI, voir (Cronin et Atkins 2000), en particulier le chapitre 14, rédigé par Jonathan R. Cole qui cite la correspondance entre Garfield et Joshua Lederberg.

5 L'institution est celle mentionnée par la revue et cette information est très lacunaire, en France et en SHS en particulier. D'où la volonté de n'autoriser que deux institutions de rattachement par chercheur affichée dans le projet « Normadresse » de l'Observatoire des sciences et des techniques.

la liste des revues qui citent ARSS (*Actes de la recherche en sciences sociales* « Euvres complètes », et « Condition Lit Double », sic...).

Ces informations permettent de déterminer le « facteur d'impact » du périodique (« Journal Impact Factor »), calculé à partir du quotient : nombre de fois où, en 2007, les articles du périodique X publiés en 2005 et 2006 sont cités dans un ensemble de périodiques de la catégorie, rapporté au nombre d'articles publiés en 2005 et 2006 dans le périodique X. On peut aussi regarder les choses sous un autre angle (*Citing*) et regarder dans quelles revues sont publiés les articles cités dans les articles d'un périodique donné pour une année de référence.

Un premier coup d'œil à cette base montre qu'elle analyse 96 revues sous la rubrique « sociologie », 57 sous « Sciences sociales et interdisciplinaires » et que, pour être précis, il faudrait aussi considérer d'autres catégories. Ainsi la *Revue française de sociologie* est répertoriée dans la première catégorie, mais *Actes de la Recherche en sciences sociales* dans la seconde. L'immense majorité de ces revues sont américaines, il y a quelques grandes revues françaises et allemandes, mais peu d'italiennes ou d'espagnoles. En gros, tout ce qui n'est pas anglophone est largement sous-représenté et souffre de bugs évidents (les umlaut et autres accents aigus, c cédille, etc. sont royalement ignorés).

Malgré tout, c'est un outil fascinant d'exploration des sous-disciplines qui, correctement manié, s'offre à nous.

Il existe d'autres outils concurrents : Scopus, Citeseer, DBLP.

Le plus connu car il est immédiatement disponible et gratuit est sans doute « Google scholar » qui, scruté avec le logiciel libre et bien nommé, « Publish or Perish » permet également de repérer les articles d'un auteur, et le nombre de citations attachées à chaque publication, et donc son « h-index ».

Le nombre d'items repéré par Google scholar est infiniment supérieur à celui du WoS, sans qu'on puisse cerner la part des doublons ou des erreurs autrement que manuellement. D'un côté on sait que le WoS ne repère qu'une fraction de la production scientifique, de l'autre on ne sait trop ce que récolte Google Scholar. Le résultat, c'est évidemment que le h-index du WoS et celui de Google scholar n'ont pas grand rapport. Nous y reviendrons.

II. Trois types d'objets à évaluer : institutions, périodiques, individus.

1. L'évaluation des institutions et le classement de Shangai.

On avait vu fleurir un classement, sur le mode du « hit parade » des lycées et des hôpitaux. Il n'y avait aucune raison que les universités échappent à ce mouvement. Malgré tout, un vent de panique s'est levé en France en 2003 lorsque le classement, établi par deux chinois de Jiao Tong University à Shangai, fait apparaître que les universités françaises ne se portent pas très bien dans la compétition mondiale. Le projet des chinois était sans malice : où devons-nous envoyer nos jeunes étudiants lorsqu'ils veulent aller étudier à l'étranger ?

Les résultats sont sans surprise : les universités qui arrivent dans le peloton de tête sont américaines ou britanniques. Néanmoins, les méthodes utilisées pour construire le classement sont discutables : observons les plus en détail.

Cinq critères sont pris en compte :

La **qualité de l'éducation** compte pour 10% et est mesurée par le nombre d'anciens élèves qui ont reçu un Prix Nobel ou une médaille Fields. (« *Alumni* »

La **qualité du corps enseignant** (40%) se compte pour moitié par le nombre d'enseignants ayant reçu un Nobel ou une médaille Fields (*Award*), pour moitié par le nombre de chercheurs fortement cités dans le Web of Science dans 21 disciplines (*HiCi*).

La **production scientifique** (40%) se mesure pour moitié par le nombre d'articles publiés dans *Nature* et dans *Science (N&S)* et pour moitié par le nombre d'articles répertoriés dans le Science Citation Index et dans le Social science Citation Index (*SCI*).

La **taille de l'institution** (10%) est la somme des indicateurs précédents rapportée au nombre d'équivalents temps plein d'enseignants dans l'université concernée (*Size*). Si ce nombre n'est pas disponible, on s'en passe en utilisant simplement les cinq indicateurs précédents, pondérés.

Par construction, pour chaque indicateur, l'institution qui reçoit la note la plus élevée est notée à 100 et les institutions suivantes voient leur score calculé comme un pourcentage de la note « record ».

Tout cela donne un outil assez joli, qui place année après année les grandes universités américaines en tête de liste, ce qui n'est pas une surprise, et les universités françaises dans une position plus que modeste, ce qui n'en est pas une non plus.

Le rang modeste tenu par les universités françaises dans ce classement s'explique assez bien si l'on considère la structure institutionnelle de l'enseignement supérieur de notre pays. Tout d'abord la balkanisation des universités consécutives à Mai 68 est fortement défavorable à la France si l'on suit les critères proposés par les chercheurs chinois. Deuxièmement, un prix Nobel français appartient bien souvent à un laboratoire rattaché à plusieurs institutions : une université et l'Ecole normale, une université et le CNRS, etc. Dès lors le « poids » de ce prix Nobel revient pour moitié à l'université en question et se perd pour l'autre moitié, puisque le CNRS n'est pas une université.

Faut-il pour autant décider de bouleverser totalement le système français d'enseignement et de recherche pour faire regrimer la France dans un classement dont les auteurs eux-mêmes reconnaissent qu'il repose sur des critères discutables ?

Ces indicateurs relèvent en effet de la méthode du réverbère : on cherche la pièce de monnaie sous le réverbère, non pas parce qu'on pense qu'elle est tombée à cet endroit, mais parce qu'il y a de la lumière. Les indicateurs proposés sont rustiques, mais, nous disent les auteurs, ils avaient le mérite d'être disponibles⁶. Dont acte.

On peut être surpris par la pratique qui consiste à traiter les universités comme les équipes de rugby. La supériorité des secondes sur les premières est qu'il y a des matches et que, normalement un match est gagné ou perdu. Il faut garder en tête qu'une évaluation n'a de sens que par rapport à un projet. Le sens du classement des universités c'est de fournir un repérage des départements où la qualité de l'enseignement est satisfaisante dans une discipline ou une sous-discipline. Ce projet semble difficilement compatible avec un classement mondial.

Reste à expliquer pourquoi le classement de Shangaï a produit en France un tel émoi chez les politiques.

Depuis, des classements concurrents sont apparus : ainsi, des chercheurs de l'Ecole des Mines ont proposé un classement alternatif⁷. Ils utilisent un critère unique, « non déclaratif et

⁶Pour plus de détails sur la construction de ce classement, voir (LIU et CHENG, 2005). Pour une critique détaillée, voir (VAN RAAN, 2005).

⁷<http://www.mines-paristech.fr/Actualites/PR/defclassementEMP.pdf>

vérifiable » : le nombre d'anciens élèves occupant un poste de numéro 1 exécutif (chief executive officer ou équivalent) dans une des 500 plus grandes entreprises internationales telles qu'elles sont repérées dans le classement des « Fortune Global 500 ». Ils montrent ainsi que la France, en dépit de son mauvais score précédent, fournit une large proportion des grands responsables sur la scène internationale privée.

Ce n'est pas le lieu de rentrer dans le détail de ces classements : on peut en particulier se demander si le fait d'avoir produit un prix Nobel est un bon indicateur de la qualité des enseignements prodigués. On aura simplement retenu que dans le classement de Shanghai, le poids des publications telles qu'elles sont repérées dans le WoS et dans le « Social Science Citation Index » est lourd. Nous verrons plus loin que c'est là une limite grave à la pertinence de l'évaluation.

Voir annexe :

Tableau Shanghai : Classement mondial des universités les 12 premières.

France : les universités françaises les mieux classées.

2. Le classement des périodiques.

Séparer le bon grain de l'ivraie est une obsession récurrente des directions scientifiques depuis un certain nombre d'années. Plus le classement est fait par des chercheurs spécialisés dans le champ qu'ils sont censés évaluer et mieux ils connaissent le sujet et sont eux-mêmes utilisateurs des périodiques, moins la démarche est toxique. Nous savons tous quelles sont, dans notre domaine, les revues que nous surveillons, où nous pouvons espérer trouver de l'information importante et fiable. Même si l'évaluation est conduite par de bons professionnels, certains problèmes demeurent.

Nous savons que ces revues ne représentent qu'une part de l'information scientifique pertinente : nous surveillons également, outre les livres, les thèses, les « working papers », les rapports, les lettres d'information et, aujourd'hui nombre d'informations scientifiques « fraîches » passent par les communications plus ou moins informelles : courrier électronique, journées d'études, sites internet, etc.

De plus, si nous regardons la base la plus « officielle », le Social Science Citation Index⁸, nous constatons que les revues que nous considérons comme majeures sont en fait noyées dans la masse.

Si l'on regarde les choses au niveau le plus général, en comparant les sciences dures aux sciences sociales, puis l'économie à la sociologie, puis enfin la sociologie anglophone par rapport à la sociologie francophone, on observe un curieux phénomène d'entonnoir.

⁸ D'autres classements concurrents apparaissent voir le « RedJasper Center for Journal Ranking », créé par des chinois. <http://www.journal-ranking.com/>.

Considérons ce tableau.

CATEGORIE	Nombre citations	Nombre Périodiques	Nombre Articles	Citations/ Articles
<i>Biologie cellulaire</i>	1199167	156	21218	56,5
<i>Astronomie et astrophysique</i>	539716	48	13810	39,1
<i>Neurosciences</i>	1246683	211	28434	43,8
<i>Economie</i>	207952	191	9245	22,5
<i>Sociologie</i>	67100	96	3099	21,7
<i>Sc. Sociales, interdisciplinaire</i>	31059	57	2325	13,4
Source : Journal citation Report				
Web of Science : année 2007				

Il existe entre les disciplines des dissymétries considérables concernant la morphologie des articles. Le nombre annuel d'articles dans les sous-disciplines de sciences « dures » se compte par dizaine de milliers et les citations dépassent le million. En économie, le nombre annuel d'articles atteint péniblement dix mille, tandis que la sociologie au sens large (sociologie + sciences sociales interdisciplinaires) en compte la moitié.

Vu à travers le prisme du WoS, le nombre d'articles, comme de citations, apparaît infiniment plus restreint dans les rubriques sociologie et sciences sociales que dans toutes les autres disciplines. Le facteur global d'impact moyen des périodiques est également plus faible en sociologie qu'en économie (0,911 en économie, 0,794 en sociologie). Si l'on confronte les ratios nombre d'articles cités en 2007/nombre d'articles publiés en 2005 et 2006 pour quelques revues de sociologie française et américaines, on comprend assez vite que les ordres de grandeur ne sont pas les mêmes.

La même observation s'impose quand l'on compare quelques grands périodiques de la discipline français et américains. Les nombres de citations et donc les facteurs d'impact sont microscopiques dans le cas des revues françaises de sociologie, même si l'on se limite aux plus connues.

Source : *Web of Science, Journal Citation Report*. Cette faible taille des indices les rend impropres à une évaluation quantitative.

Cependant, l'ambition de classer les revues de sciences sociales est réapparue récemment, d'abord au CNRS en 2004, puis plus récemment à La Fondation européenne de la science (ESF) et, en France, à l'AERES. Ces classements de revues ont provoqué chaque fois une vague de protestation. Quelles en sont les raisons ?

Tout d'abord, la question de la légitimité des évaluateurs est posée.

On ne sait pas qui sont les « experts » ayant effectué ce classement. On attend en effet une sorte d'impartialité des experts qui devraient représenter un panel des différentes branches et tendances idéologiques et scientifiques présentes dans la discipline.

. Ensuite, le contenu et le choix des classements est mis en cause. Il y a des oublis, des critiques de la distinction entre revue de rang A, B et C.

Le classement concernant les revues étrangères est remis en cause par beaucoup.

D'autres enfin mettent radicalement en cause la légitimité d'un tel classement.

A cet égard, de bons arguments ont été mis en avant dans une note émanant de la rédaction de la revue *Actes de la recherche*⁹. On ne saurait, disent les auteurs, construire un axe allant des « bonnes » revues qui méritent d'être financées à de « mauvaises » revues qui ne le méritent pas. Certaines revues correspondent à des thématiques très pointues et ont leur utilité même si leur niveau de diffusion est restreint. En outre, une revue n'est pas une simple machine à produire des articles, c'est un levier d'animation de la communauté scientifique. Les revues doivent pouvoir naître, vivre, disparaître : ce sont aux responsables des revues et à leurs lecteurs d'apprécier la situation. Les financeurs, eux, doivent prendre leurs décisions au cas par cas, non au vu d'un quelconque facteur d'impact.

Plus que tout, on s'interroge sur l'usage qui doit être fait d'un tel classement. Ce mélange d'opacité et d'information imparfaite génère dans le milieu une vive inquiétude.

Un représentant de l'AERES déclare que ce classement n'a pas été dressé à des fins pratiques d'évaluation des personnes ni pour modifier les dotations financières des revues, mais qu'il s'agit de "délimiter le périmètre scientifique des revues". A une autre occasion, le même représentant affirme qu'il s'agit d'une opération pédagogique incitant les chercheurs à se placer « dans la compétition internationale ».

Il est clair que certains espèrent par un classement des revues faire l'économie d'une évaluation des articles publiés par les chercheurs : un article publié dans une revue de rang A entraînerait automatiquement la confiance des évaluateurs.

En fait, on trouve de bons articles dans des revues qui ne sont pas de rang B ou C et des articles médiocres dans des revues de rang A.

Les comités de rédaction fonctionnent eux aussi au moins partiellement à la réputation. L'argument "on ne peut pas refuser un article au Professeur X" est parfois mis en avant et entendu. Les revues sont elles aussi des systèmes sociaux soumis à des luttes d'influence et le texte de l'article examiné est également évalué à travers le prisme de ce que l'on sait de l'auteur et de son entourage, même si les papiers sont souvent "anonymisés" pour les besoins de l'évaluation.

Ceci ne doit pas être entendu comme une critique du mode de fonctionnement des

⁹Actes de la recherche en sciences sociales (2004), Pour une Realpolitik de la recherche, Au sujet de l'enquête du CNRS sur les périodiques aidés par le département des sciences de l'homme et de la société (SHS).

revues, mais comme une mise en garde contre une conception "chosiste" des publications. Les revues ne sont pas de simples machines à publier des textes. Elles ont un rôle de stimulation de la vie intellectuelle : initier de nouveaux travaux en commandant des articles et des numéros spéciaux, par exemple.

En fait, la publication des listes de classement des revues éveille une inquiétude profonde et fondée dans la communauté scientifique parce qu'il y a raison de penser que ce classement des revues est dirigé vers une évaluation des personnes.

En effet, une notion est apparue récemment, celle de « chercheur publiant ». Cette notion recouvre deux arrières-pensées.

La première concerne les enseignants du supérieur, Maîtres de conférence et Professeurs, qui, en plus de leur charge d'enseignement parfois lourdes et des tâches administratives multiples et grandissantes, doivent assurer une production scientifique. Pour ceux-ci, la barre a été fixée à au moins deux publications dans une revue de rang A sur une période de 4 ans.

La seconde concerne les chercheurs du CNRS pour lesquels, la recherche étant l'activité principale, la barre a été définie à 4 articles sur la même période. Jadis, les rapports de recherche se mesuraient au poids, aujourd'hui on pèse les chercheurs.

On nous rétorquera : un article tous les deux ans (ou tous les ans), ce n'est pas très exigeant.

Il a aussi été suggéré que, tout compte fait, les publications de rang B comptaient aussi. Dès lors, pourquoi deux catégories et non une seule ? Il a aussi été suggéré que les livres seraient pris en compte dans l'évaluation (mais quelle serait la métrique : deux noires valent une blanche, deux articles valent un livre ??).

Peu important les détails : ce qui compte c'est que le classement des revues est devenu un outil pour faire la chasse au chercheur « paresseux », celui qui n'est pas un « chercheur publiant ». Or au CNRS comme dans l'enseignement supérieur, le chercheur « paresseux » est en fait une espèce extrêmement rare. Pour lutter contre cette espèce rare, on crée des dispositifs pour lesquels le remède est pire que le mal. Malgré les démentis de toute sorte, on est obligé de voir un lien entre évaluation des revues et évaluation des personnes.

Abordons donc maintenant l'évaluation des personnes.

3. L'évaluation des personnes.

1. Le h-index, sa construction et ses possibles effets pervers.

L'outil rêvé de l'évaluation quantitative a été produit récemment par un physicien de l'Université de Californie à San Diego¹⁰. Cet indice est calculé, pour un chercheur donné, en comptant le nombre h d'articles cités plus de h fois. Si j'ai écrit douze articles, et que cinq ont été cités plus de 5 fois, mon h-index est 5. Rustique, mais efficace. L'auteur soutient que l'impact scientifique global de deux chercheurs sera identique si leur h-index est identique même si le nombre d'articles qu'ils ont publiés est différent. Le mérite de cet indicateur, nous dit son auteur, est d'être calculable aisément à partir de la base du Web of Science.

Ce qui est plus intéressant, ce sont les restrictions que l'auteur apporte lui-même quant à la validité de son indicateur. Tout d'abord, il souligne qu'un seul indicateur ne saurait suffire à mesurer la qualité d'une production scientifique. Ensuite, s'il affirme que, de deux chercheurs (ayant même ancienneté dans le monde scientifique) ayant des indices différents, celui qui a l'indice le plus élevé est susceptible d'être le chercheur le plus accompli, il affirme également que la réciproque n'est pas nécessairement vraie : autrement dit, on peut trouver des chercheurs ayant mis à jour des résultats scientifiques importants qui ont un indice peu élevé. Ainsi, un chercheur ayant produit un nombre limité d'articles très fréquemment cités aura un h-index petit. C'est le cas, en particulier, nous dit-il, pour des chercheurs qui travaillent sur des domaines de recherche à faible visibilité. (« non mainstream »).

Enfin, il est évident que le sens de l'indicateur à des niveaux élevés (il cite des h-index de 90 ou plus pour des chercheurs célèbres ayant atteint la maturité) n'est pas le même que lorsque l'indicateur se situe à des niveaux très bas et avec une amplitude faible à l'intérieur d'une discipline.

Pour explorer cette voie, nous avons pris quelques exemples de sociologues seniors français et américains et nous avons comparé leur h-index en utilisant deux bases différentes : celle du WoS et Google Scholar couplé avec le logiciel libre « Publish or Perish ». Le choix de ces exemples est arbitraire et j'espère que mes collègues ne m'en voudront pas d'avoir voulu montrer que, même des chercheurs français de bonne réputation ont, sauf exception, des h-index relativement petits.

¹⁰ Voir Hirsch J.E. (2005), An Index to quantify an individual's scientific output, *Proceedings of the National Academy of Sciences*, vol. 102, n. 46, pp. 16569-16572.

Dans ce tableau sont sélectionnés un certain nombre de sociologues seniors anglo-saxons et français. On remarquera que les francophones ont un nombre de publications comparables à celui des américains. Ce résultat est surprenant, car plusieurs revues importantes dans la discipline (*L'Année sociologique*, *Sociétés contemporaines*, *Genèses*, etc..) sont absentes de la base du WoS. En revanche, les francophones sont beaucoup moins fréquemment cités, sauf s'ils ont été traduits, ce qui constitue des exceptions rares.

On doit signaler ici un effet pervers possible de l'usage de la bibliométrie : elle peut entraîner, de la part de chercheurs soucieux de leur carrière et donc anxieux d'être cités, un recul massif de la langue française dans des domaines où cela n'est ni souhaitable ni inévitable. En effet, il est clair que les chercheurs francophones qui ont un h-index « raisonnable » sont tous des chercheurs qui ont une stratégie internationale et qui ont publié dans des revues anglo-saxonnes.

On remarquera également que le « h-index » calculé à partir de Google est généralement plus élevé que celui qui émane du WoS, mais le rapport entre les deux indicateurs est des plus inconstant puisqu'il varie entre 0,8 et 15. Un joli exemple de ce « grand écart » est fourni par l'article de J.S. Coleman (1988), *The Social Capital in the creation of Human Capital*, publié dans *l'American Journal of Sociology*, cité 637 fois selon le Web of Science, mais 8344 fois selon Google Scholar.

Calculé à partir du WoS ou de Google, Cet indicateur est le plus souvent si petit pour les francophones et varie de façon si erratique qu'on peut légitimement se demander si on peut en tirer quoi que ce soit. Faut-il penser qu'entre un chercheur dont le h-index est 1 et celui dont le h-index est 3 il existe une différence de qualité incontestable ?

A ce stade, il nous faut revenir sur une notion évoquée rapidement au début de cet article, la citation.

2. *Que signifie citer ?*

Sur le sujet, il existe une abondante littérature. Le credo initial, formulé par les sociologues et historiens des sciences américains, est ainsi formulé par l'un d'entre eux,

« Extensive past research indicates that citations are a valid indicator of the relative quality or impact of work. The number of citations is highly correlated with all other measures of quality that sociologists of science have employed. As long as we keep in mind that research of high quality is being defined as research that other scientists find useful in their current work, citations provide a satisfactory indicator. Citations do not measure the absolute or « objective quality of research but they do measure the currently assessed value of work to colleagues who are themselves doing research. (Cole, 1983). »

« Une masse de travaux antérieurs montre que les citations sont un bon indicateur de la qualité relative ou de l'influence d'une oeuvre. Le nombre de citations est fortement corrélé avec toute autre mesure de la qualité employée en sociologie des sciences. A condition de ne pas perdre de vue qu'une recherche de qualité est définie comme celle que d'autres chercheurs trouvent utile à leur démarche scientifique présente, les citations peuvent être considérées comme des indicateurs satisfaisants. Les citations ne donnent pas une mesure de la qualité "absolue" ou "objective" de la recherche, mais elles reflètent le jugement porté par des collègues qui conduisent eux-mêmes des recherches. » (Cole, 1983).

Chez cet ardent défenseur de la bibliométrie, on trouve déjà une réserve importante : le comptage des citations ne saurait refléter la qualité intrinsèque des travaux scientifiques.

On peut faire un pas de plus en cherchant ce qui se cache derrière les citations : pourquoi cite-t-on les travaux scientifiques antérieurs? Les sociologues des sciences, Robert K. Merton en tête, nous enseignent que le monde de la recherche ne fonctionne pas comme un marché: c'est la publication d'une découverte ou d'une idée qui en confère à son auteur la paternité ou la propriété. La citation de l'auteur d'une publication est donc l'acte par lequel on rend hommage au père de l'invention ou de l'idée. Dans la morale du milieu, reprendre une idée nouvelle sans mentionner à qui on l'« emprunte », c'est du vol. La citation est un procédé de reconnaissance de dette. A ce titre, le repérage des citations devrait permettre de reconstituer un certain nombre de processus de filiation intellectuelle.

La citation est aussi beaucoup d'autres choses : allusion rhétorique destinée à rassurer le lecteur et à le persuader du bien fondé et de l'intérêt de l'article, geste qui se veut élégant pour montrer que l'on connaît la littérature sur le sujet, etc. Un auteur connu a plus de chances d'être cité que son co-auteur, un peu moins connu.

Il y a aussi la citation négative : si je cite Sokal, ce n'est probablement pas parce que je me réclame de lui. Il y a donc au moins deux faces à la citation : elle est à la fois signe de

reconnaissance et d'allégeance intellectuelle et outil de persuasion, procédé rhétorique. Le malheur, c'est que distinguer une face de l'autre n'est pas possible de façon automatique (Cozzens, 1989) Je ne puis dans le cadre de cet article traiter en détail d'un sujet qui a été maintes fois abordé ailleurs. (Kaplan, 1965 ; MacRoberts and MacRoberts, 1986; Zuckerman 1987; Gilbert, 1977). L'influence telle qu'elle est vue à travers les références, subit des processus de déformation si nombreux que toute utilisation purement automatisée, même aux seules fins de la sociologie des sciences, semble interdite.

« The mere presence of a reference is not a marker of influence, nor is the absence of reference evidence that it is uninfluential. In citation analysis, if it is to be taken seriously, investigators must first descend to the documents from which these data are derived in order to reconstruct influences before proceeding further. » (MacRoberts and MacRoberts, 1986).

“La simple présence d'une référence ne saurait être considérée comme un signe d'influence, non plus que son absence comme un signe de non influence. Si l'on veut faire de l'analyse des citations une entreprise sérieuse, il convient, avant toute autre chose, de revenir d'abord aux documents de première main afin de reconstruire les influences” (MacRoberts et MacRoberts, 1986).

A fortiori, s'il s'agit de l'évaluation des personnes, des précautions encore plus lourdes s'imposent.

En tout état de cause, il faut se garder d'oublier que les articles cités sont le résultat de filtrages successifs ; ils ne sont que le dernier niveau, étroit, d'une fusée à plusieurs étages. Les chercheurs produisent d'abord des communications ou des rapports, puis rédigent des articles, qu'ils soumettent à des revues ; une fraction variable de ces manuscrits arrive à publication, un sur dix ou un sur vingt, parfois moins. Enfin, un article publié peut, par une série de processus sociaux que l'on connaît fort mal, parvenir à être cité. La citation est une mesure raisonnable de la visibilité d'un résultat scientifique. Quand bien même l'on voudrait départager les chercheurs qui travaillent de ceux qui « ne travaillent pas » (à supposer que cette expression ait un sens), le critère de citation est une mesure bien rustique. Sous des parures d'objectivité, cet indicateur réintroduit en fait ce qu'il prétendait éliminer ; c'est-à-dire, le jugement par les pairs.

III. Remarques conclusives : au delà des sciences sociales et de la France

Jusqu'à présent j'ai choisi d'examiner le problème dans un cadre étroit, celui des sciences sociales et plus spécifiquement de la sociologie française. Il est vrai que l'utilisation des méthodes quantitatives à des fins d'évaluation y est particulièrement délicate, notamment pour deux raisons, l'importance jouée par les livres dans ce domaine et la prégnance de la langue française qui cantonne la discipline dans un espace d'échanges restreints.

Cependant, les spécialistes de sciences sociales ont trop tendance à considérer qu'ils constituent un cas à part dans la communauté scientifique. La confrontation avec d'autres disciplines pour envisager si leur perception de la bibliométrie est moins critique que la notre peut s'avérer utile.

Je voudrais maintenant montrer que cette critique déborde largement le cadre français et ne se limite pas au champ de la sociologie.

1. Dans les sciences « dures ».

On découvre assez rapidement que les réticences à l'égard de la bibliométrie comme outil d'évaluation ne concernent pas principalement les Sciences de l'Homme et de la Société.

On trouve des analyses critiques sous la plume de chercheurs de disciplines diverses. Citons quelques exemples.

a) Biologie.

Peter Lawrence, chercheur en Biologie à l'université de Cambridge tourne en dérision les 48 citations dont l'un de ses articles est l'objet : « seulement 8 se réfèrent correctement à ce que j'ai écrit. Les 40 autres sont soit de pures allusions, soit erronées. »

Il souligne la dégradation que la manie de la mesure a fait subir à l'activité scientifique ; en vingt ans, on est passé d'une activité consacrée à la poursuite de la découverte, à une activité centrée sur la nécessité de rédiger des articles et de les faire publier dans le meilleur périodique possible. Les ravages de cette pratique sont multiples :

-Multiplication des comportements amoraux, car les cas où un chercheur va être amené à signer un article auquel il n'a que très faiblement contribué se multiplient.

-« Saucissonnage » des sujets afin de produire un nombre d'articles le plus élevé possible avec un seul résultat.

-Aversion au risque, attraction pour les sujets rebattus et évitement des sujets nouveaux, considérés comme plus risqués et moins susceptibles d'être évalués positivement.

-Multiplication des comportements carnassiers et effacement de chercheurs de valeur plus timides. Or, nous dit, Lawrence, « on ne tient pas la preuve que des chercheurs plus combattifs soient plus créatifs que les autres ».

b) Sciences de la terre et de l'univers.

Au CNRS dans la section compétente du comité national, des expériences ont été conduites par Yves Langevin, Président de section. Dans cette commission, les indicateurs bibliométriques ont été utilisés, aux côtés d'autres critères, pour le concours DR2.

Une batterie d'indicateurs extrêmement complexes a été construite: pas moins de 14 indicateurs bibliométriques ont été appliqués aux candidats à la promotion de Directeur de recherches. Les résultats ont ensuite été confrontés au jugement de promotion final.

Si cette batterie d'indicateurs permet, en gros, de confirmer si les candidats se trouvent ou non dans la première moitié du classement final, en revanche la corrélation avec les positions dans les premiers rangs n'est pas bonne. D'autres indicateurs (jugements par les pairs, positions de responsabilité dans le milieu scientifique, etc.) sont de poids quand il s'agit de déterminer les positions dans les premiers rangs.

On notera que dans cette section, toutes les revues scientifiques sont considérées comme étant de rang A.

Cette expérience montre que l'utilisation de la bibliométrie pour l'évaluation pénalise les chercheurs qui ont une mobilité thématique (puisque'il faut plusieurs années pour se refaire une réputation à l'intérieur d'un groupe thématique donné).

Enfin, ici comme ailleurs, on vérifie que ce système entraîne une série d'effets pervers dont l'uniformisation des pratiques scientifiques, la multiplication des publications avec tronçonnement des résultats, etc. Le Président Langevin étant un optimiste, il prédit qu'à terme, ce système d'évaluation par la bibliométrie s'autodétruit, par uniformisation des pratiques.

c) Mathématiques.

La communauté des mathématiciens est, on le sait, particulièrement bien organisée et fortement internationale. L'Union internationale des mathématiciens (IMU), en coopération avec l'ICIAM (International Council for Industrial and Applied Mathematics) et l'IMS

(Institute of Mathematical Statistics) a publié un rapport assez dévastateur sur l'utilisation de la bibliométrie pour l'évaluation.

L'exemple de cette discipline est tout a fait surprenant car on pouvait s'attendre à ce que les publications de mathématiques, discipline noble par excellence, soient mieux couvertes que d'autres dans les bases documentaires. Il n'en est rien: ce texte souligne le fait que les mathématiques ont une culture bien particulière de la citation, que moins d'un article de mathématiques sur deux est recensé dans le Web of Science. Le rapport insiste sur le caractère instable du facteur d'impact, son manque de finesse dans la mesure ; il rappelle qu'en aucun cas les indicateurs quantitatifs ne doivent être manipulés indépendamment d'autres critères.

Plus intéressant encore, ce rapport souligne l'absence de théorie sous-jacente à toutes ces méthodes quantitatives : on suppose simplement qu'un facteur d'impact élevé dénoterait une bonne qualité de la revue ; il n'y a pas une théorie permettant de fonder cette supposition selon laquelle l'un et l'autre seraient corrélés.

d) Informatique.

La commission d'évaluation de l'INRIA a produit sur le sujet un rapport fort intéressant (Merlet, Robert, Segoufin, 2007). On peut y lire en conclusion : « *Si les indicateurs peuvent donner des tendances sur un nombre réduit d'aspects de la vie scientifique, il convient d'être très circonspect dans leur usage en raison de la possibilité d'interprétations erronées, des erreurs de mesure (souvent considérables) et des biais dont ils sont affectés. Un usage abusif des indicateurs est facilité par la nature chiffrée du résultat qui introduit la possibilité d'établir dans l'urgence toutes sortes de statistiques, sans se préoccuper d'en analyser la qualité et le contenu, et en occultant l'examen d'autres éléments de la vie scientifique comme, par exemple, l'innovation et le transfert intellectuel et industriel.* »

Suivent un certain nombre de recommandations qui invitent à la plus grande prudence dans le maniement d'informations dont le caractère lacunaire et entaché d'erreurs est constamment souligné. Ces indicateurs bibliométriques peuvent éventuellement servir à dégager « le haut de la pyramide », mais le jugement par les pairs reste l'outil le plus sûr et, paradoxalement, le moins subjectif...

2. Au delà des frontières.

Royaume-Uni, Australie, Etats-Unis, Canada, Brésil la liste est longue des pays qui, aux côtés de la France et souvent avant elle, sont touchés par le virus de l'évaluation. La création d'une société où l'ordinateur remplacerait l'œil humain pour évaluer les productions de la recherche semble être une ambition assez répandue.

Un rapport récent annonce qu'au Royaume-Uni, les méthodes bibliométriques sont désormais appelées à remplacer les méthodes traditionnelles d'évaluation par les pairs.

« *Metrics rather than peer-review will be the focus of new system and it is expected that bibliometrics will be the central quality index in this system.* (Evidence Report (2007)).

Trop nombreux sont les chercheurs qui dénoncent ce penchant pour les indicateurs pseudo-quantitatifs , tout en considérant trop souvent qu'il s'agit là d'une fatalité à laquelle on ne peut échapper.

Il est urgent pour la communauté scientifique des sociologues de s'informer et d'élaborer une réflexion collective sur le sujet : que nous le voulions ou non, pour de bonnes ou de mauvaises raisons, ces outils bibliométriques sont déjà utilisés¹¹. Certains de ces usages

11

peuvent être relativement raisonnables, mais ils ne sont ni neutres ni sans effet de retour sur les pratiques de recherche. D'autres usages de ces indicateurs sont illégitimes. Il reste à obtenir que des outils qui ont quelque intérêt ne soient pas utilisés en dépit du sens commun. Les grandes manœuvres actuelles autour de la bibliométrie me rappellent étrangement les querelles passées autour de la mesure de l'intelligence par le « Quotient intellectuel » : mesurer rassure, mais l'on ne mesure pas ce que l'on croyait ou voulait mesurer.

Rappelons que Hirsch, l'inventeur du h-index avait bien souligné que cet indicateur n'a pas de sens pour des chercheurs dont l'ancienneté professionnelle est faible. En de multiples occasions, des figures éminentes de la sociologie des sciences ont affirmé que les scores bibliométriques ne devaient en aucune façon servir à évaluer les personnes (Merton, 1979; Cole, 2000). Ils ont insisté sur la nécessité d'appliquer ces méthodes à des agrégats statistiques, et de se garder d'affirmer qu'un chercheur dont l'article a été cité quinze fois a plus d'impact que celui qui a été cité dix fois (Cole, 2000).

Aucun de ces messages ne semble avoir été entendu. On utilise aujourd'hui l'outil bibliométrique en dehors du cadre pour lequel il a été conçu. Pourtant, comme disait Eric de Dampierre, un piano n'est pas fait pour être joué avec les pieds. La mesure du nombre de citations ne saurait se substituer à l'évaluation par les pairs. Confondre la mesure de la productivité avec celle de la créativité scientifique est une erreur grave. Peut-on dire que Leonardo da Vinci est un peintre de faible envergure à raison du nombre restreint de ses toiles ?

Les outils bibliométriques sont précieux pour explorer la structure d'un domaine de recherche. En revanche, le recours à des indicateurs quantitatifs ne permet pas de jauger l'importance d'une idée ou d'un résultat scientifique. Seule la lecture de la production scientifique ou au moins d'échantillons de celle-ci permet d'apprécier la solidité et la fécondité d'une œuvre. Seuls des

savants compétents spécialistes du domaine ou d'un domaine voisin sont à même de porter un jugement raisonnable et fondé. Les indicateurs bibliométriques, pour rassurants qu'ils soient, ne peuvent que contribuer à détériorer la qualité de la recherche en détournant les chercheurs des sujets nouveaux et complexes et en les encourageant à se concentrer sur les sujets plus faciles

et moins féconds. Avec tous les défauts que nous lui connaissons, le système du peer review est le moins nocif. A nous de mettre en place les contre-pouvoirs permettant d'en assurer un usage impartial.

REFERENCES

- Cole J. et Cole S. (1971) Measuring the Quality of Sociological Research : problems in the use of the Social Sciences Citation Index, *The American Sociologist*, vol. 6, (February), 23-29.
- Cole J. R. (1983) The Hierarchy of Science?, *American Journal of Sociology*, vol. 89, n°1, pp 111-139.
- Cole J.R. (2000) A short History of the Use of Citations as a Measure of the Impact of Scientific and scholarly work, in Cronin B. and Atkins H.B., ed. *The Web of Knowledge*, a Festschrift in Honor of Eugene Garfield, pp. 281-300., Assis Monograph Series, <http://www.columbia.edu/cu/univprof/jco/publications.html>
- Cozzens S.E. (1989) What do citations count?, *Scientometrics*, vol. 15, n. 5-6, pp. 437-447.
- Cozzens S.E., Healey, P., Rip A., Ziman J. (1990), *The Research System in Transition*, Kluwer.
- Evidence Report (2007), The use of bibliometrics to measure research quality in the UK higher education system. <http://bookshop.universitiesuk.ac.uk/downloads/bibliometrics.pdf>
- Gilbert G.N. (1977), Referencing as Persuasion, *Social Studies of Science*, Vol. 7, 113-122.
- Gingras Y. (2008), La fièvre de l'évaluation de la recherche. Du mauvais usage de faux indicateurs. CIRST, UQAM. <http://www.cirst.uqam.ca>.
- Hirsch J.E. (2005), An Index to quantify an individual's scientific output, *Proceedings of the National Academy of Sciences*, vol. 102, n. 46, pp. 16569-16572.
- Lawrence P., (2007) The mismeasurement of Science, *Current Biology*, Vol. 17 n°15., pp 583-585.
- Liu N.C. et Cheng Y. (2005) Academic Ranking of World Universities- Methodologies and Problems, *Higher Education in Europe*, Vol. 30, N°2.
- Liu N.C., Cheng Y. Liu L. ((2005) Academic Ranking of World Universities using Scientometrics – A comment on Fatal Attraction, *Scientometrics*, vol. 64, N°1, pp 101-109.
- MacRoberts M.H. et MacRoberts B.R., (1986) Quantitative Measures of Communication in Science : a study of the Formal Level, *Social Studies of Science*, Feb., Vol. 16 n°1, pp 151-172.
- Merton R. K. (1979). Préface à E. Garfield, *Citation Indexing, Its Theory and Application in Science, Technology and Humanities*. Wiley. <http://www.garfield.library.upenn.edu/cifwd.html>
- Merlet J.P., Robert P. Segoufin L., (2007) Que mesurent les indicateurs bibliométriques ?, Document d'analyse de la commission d'évaluation de l'INRIA. <http://www.inria.fr/inria/organigramme/ce.fr.html> puis cliquer sur « indicateurs bibliométriques ».

Van Raan A.F., (2005) Fatal Attraction: conceptual and methodological problems in the ranking of universities by bibliometric methods, *Scientometrics*, vol 62, n°1, pp 133-143.

Zuckerman H. (1987) Citation analysis and the complex problem of intellectual influence

Scientometrics, 12, n°5-6, pp.329-338. ANNEXE 1 : Le classement de Shanghai : rangs 2003-2007.

Classement de Shanghai 2007 : Institutions françaises d'enseignement supérieur.