L'évaluation en sciences exactes : quand la quantité tue la qualité

Alexandre Matzkin

L'évaluation scientifique a toujours fait partie du paysage de la recherche en sciences exactes si l'on entend par là l'existence de procédures pertinentes d'examen par les pairs – qu'il s'agisse de la soumission d'un article à une revue scientifique, du recrutement et de l'affectation thématique d'un chercheur, ou du financement de programmes de recherche nécessitant des investissements importants en moyens humains et matériels. Toutefois, la nature et les modalités de l'évaluation scientifique ont été profondément bouleversées ces dernières années par l'apparition de deux éléments connexes. D'une part, la mise en place, par les tutelles politiques, d'outils permettant un pilotage resserré de la politique scientifique. D'autre part, l'informatisation aidant, les bases de données bibliographiques se transforment progressivement en outils mesurant la qualité de la production scientifique, donnant naissance à un foisonnement de classements en tous genres : les indices de citation et les facteurs d'impact servent ainsi de base pour classer les chercheurs, les laboratoires, les universités, les revues...

Or ces classements, souvent très médiatisés, tendent à se substituer à l'évaluation scientifique. Cette évolution est d'autant plus encouragée par la technocratie et les politiques gouvernementales que les classements sont supposés fournir des chiffres « objectifs » permettant de définir l'excellence sans avoir besoin de comprendre quoi que ce soit. Pourtant, alors que cette évolution commence à être internalisée par les scientifiques eux-

Cités 37, Paris, PUF, 2009

mêmes, des sociétés savantes et des instances d'évaluation tirent la sonnette d'alarme, pointant les dangers que cette pression du quantitatif fait courir à la recherche, à sa diversité et à sa qualité.

Il n'est pas lieu de revenir ici sur la diversité des formes que peut prendre l'évaluation par les pairs, exercice collectif et contradictoire (généralement mené dans le cadre de commissions *ad hoc* ou permanentes) où les différentes facettes des activités scientifiques sont examinées. Le contexte dans lequel s'est mis en place le processus actuel est cependant révélateur. On peut ainsi remarquer l'encadrement de plus en plus resserré des procédures et des conséquences de l'évaluation par le pouvoir politique dans la plupart des pays et la tendance, depuis une vingtaine d'années, à structurer cette évaluation sur le plan national. La principale raison de cette tendance réside dans l'importance prise par la R&D (recherche et développement) dans la compétitivité économique d'un pays ; certains corollaires, comme l'émergence d'un marché globalisé de la connaissance, ou encore une vision idéologique de la finalité de la recherche et l'organisation conséquente permettant le meilleur rendement coût/bénéfice, jouent également un rôle. En France, où l'espace de la recherche est depuis longtemps structuré au niveau national au moyen d'organismes comme le CNRS, l'évaluation[1] était historiquement dévolue au Comité national de la recherche scientifique (CoNRS), un organisme placé auprès du CNRS (mais n'en faisant formellement pas partie), composé de membres élus par l'ensemble de la communauté scientifique et de membres nommés par le gouvernement. Le décret de 1982 sur l'organisation du CNRS présente explicitement le CoNRS comme *« une instance de conseil et d'évaluation »* des programmes de recherche menés dans les unités propres ou associées au CNRS, de la conjoncture scientifique, et du personnel[2] – dans les textes antérieurs, comme le décret de 1959, le terme « évaluation » ne figure pas, mais l'esprit est le même. La création en 2006 d'une agence gouvernementale d'évaluation de la recherche, l'AERES, a dépossédé le CoNRS de l'évaluation des unités *stricto sensu,* la prospective et l'évaluation du personnel demeurant l'apanage du Comité national. Sur le fond, l'apparition de l'AERES a peu modifié la situation en sciences exactes : en effet, contrairement aux sciences humaines, il n'y a quasiment

1. J. Fossey, « L'évaluation scientifique au CNRS », *Revue pour l'histoire du CNRS,* nº 8, 2003.
2. Décret nº 82-993 du 24 novembre 1982.

pas d'équipes de recherche purement universitaires (lesquelles n'étaient auparavant évaluées par aucune instance), puisque même les unités où le personnel universitaire est ultra majoritaire, comme c'est le cas général en mathématiques, sont associées au CNRS. Sauf exception, les comités de visite de l'AERES travaillent comme ceux du CoNRS, mais de manière moins approfondie (les visites étant beaucoup plus courtes), sans membres élus (afin d'en faciliter le pilotage), et en excluant la participation des ingénieurs et techniciens (pourtant plus nombreux que les chercheurs dans nombre d'unités).

Aussi, la principale inflexion dans l'évaluation en sciences exactes réside-t-elle dans la prépondérance acquise par les facteurs quantitatifs basés en majeure partie sur la bibliométrie – des facteurs qui constituent de mauvais indicateurs et sont, au demeurant, mal utilisés, pour reprendre le titre d'une étude récente[1]. L'arbitraire des classements, très médiatisés, censés définir les meilleures universités est bien connu. Il suffit, pour s'en convaincre, de comparer ces derniers entre eux[2]. Par exemple, le classement du *Times* de Londres surreprésente les universités anglaises et celles de l'ex-Empire britannique, du Canada et de l'Australie ; le classement de Leyde compte plusieurs universités hollandaises dans son « top 15 », et l'École des mines de Paris met à l'honneur les grandes écoles françaises. Pis encore, le plus connu, le classement de Shanghai, qui avantage les universités américaines, ne peut être reproduit en suivant la méthodologie indiquée pour l'établir[3]. Reste à savoir si la fascination qu'exercent ces classements dans les ministères est réelle ou feinte (afin de justifier des réformes comme en France ou en Italie).

L'utilisation des indices bibliométriques est moins connue et pourtant plus dangereuse. S'il était autrefois évident que le nombre d'articles publiés – ou ce nombre normalisé par le nombre d'auteurs – pouvait difficilement refléter la qualité de la recherche, les directions des organismes scientifiques, ainsi qu'une partie de la communauté des chercheurs, sont

1. Y. Gingras, « La fièvre de l'évaluation de la recherche : du mauvais usage de faux indicateurs », *Note de recherche du CIRST 2008-05,* Centre interuniversitaire de recherche sur la science et la technologie, Montréal, 2008.

2. Ces classements peuvent être consultés sur Internet aux adresses respectives : http://www.timeshighereducation.co.uk ; http://www.cwts.nl/cwts/LeidenRankingWebSite.html ; http://www.ensmp.fr/Actualites/PR et http://www.arwu.org.

3. R. V. Florian, « Irreproducibility of the results of the Shanghai Academic Ranking of World Universities », *Scientometrics,* 72, 2007, p. 25-32.

pourtant aujourd'hui convaincues que les indices de citations suffisent par eux-mêmes à détecter la recherche d'excellence. Ces indices sont construits à partir de bases de données collectées par des sociétés privées, principalement l'*Institute for Scientific Information* (ISI) de Thomson-Reuters et Scopus d'Elsevier, qui vendent très cher l'accès à leurs services. Les revues « excellentes » sont celles dont les articles sont cités le plus grand nombre de fois ; le facteur d'impact (IF) d'une revue pour l'année *n,* l'indice le plus utilisé, est défini par le nombre de citations d'articles parus dans la revue au cours des deux années précédentes $n-1$ et $n-2$. L'IF privilégie ainsi le très court terme, ce qui n'est pertinent que pour certaines sous-disciplines, mais encourage partout les effets de mode. L'IF varie également en fonction de la nature des articles (communications rapides, articles de revue, nombre moyen de références dans les articles, langue, etc.) et seul un argument circulaire permet d'en faire la représentation de la qualité académique d'une revue. Pourtant l'IF sert de support au classement des revues, le plus souvent en trois catégories. L'extrapolation lors de l'évaluation individuelle des chercheurs consistant à attribuer un score à un article en fonction de l'IF de la revue dans laquelle il paraît est un non-sens statistique (malheureusement répandu) ; l'IF est un agrégat et ne peut donc refléter la qualité intrinsèque d'un article.

L'indice de citations le plus utilisé lors de l'évaluation individuelle, malgré l'introduction récente (en 2005) par un physicien, est le facteur H. Ce facteur, calculé automatiquement par les bases de données, répertorie le nombre *n* d'articles d'un auteur cité au moins *n* fois. Il reprend ainsi le postulat des indices de citation : un auteur cite les articles qui l'ont influencé, et la qualité d'un article est corrélée avec son influence, donc avec le nombre de citations. Il y ajoute une mesure globale de l'influence : mieux vaut avoir publié trente articles cités plus de trente fois (H = 30) qu'un seul article cité mille fois et tous les autres moins de dix fois (H = 9). Des dizaines d'autres indices, censés pallier la simplicité du facteur H, ont été proposés, pondérant de différentes manières le nombre et l'âge des auteurs, les autocitations, la moyenne des citations de la revue ou du domaine, etc. C'est pourtant la simplicité conceptuelle du facteur H qui en a fait le succès : l'on entend souvent dans des commissions d'évaluation invoquer le facteur H et le nombre de citations des candidats à la place d'une argumentation qualitative globale. Des cas de promotion refusée pour cause d'indices de citation insuffisants, alors que l'examen qualitatif du dossier scientifique était très

positif, ont déjà été signalés. Une opération menée depuis un an par la direction du CNRS, et – provisoirement ? – enterrée suite à de nombreuses protestations, a consisté à établir pour chaque laboratoire une liste individuelle du nombre de publications et des indices de citation, affectant un score à chaque équipe et aboutissant à une note globale pour l'unité. Le service en charge de cette opération était tout à fait incapable de comprendre la moindre problématique scientifique des équipes examinées, mais il n'en demeure pas moins que le but (inavoué) était de récolter des statistiques pour caractériser les laboratoires afin de sélectionner ceux que le CNRS conserverait en cas de recentrage de l'organisme sur un petit nombre d'unités. En Grande-Bretagne, les autorités en charge de la RAE *(Research Assessment Exercise),* gigantesque opération d'évaluation ayant lieu tous les sept ans dont les résultats déterminent les financements futurs des équipes de recherche, ont décidé de baser le prochain exercice d'évaluation, prévu vers 2015, exclusivement ou essentiellement (selon les disciplines) sur des indices quantitatifs[1]. Le pilotage de la transition au quantitatif a d'ailleurs été confié à une société privée ayant des liens avec l'ISI évoqué plus haut.

Cette vogue du quantitatif s'explique parfois par la facilité, car on évite les discussions de fond ainsi que d'avoir à justifier les choix de manière argumentée. Quant aux tutelles politiques, elles reprochent à l'évaluation qualitative par les pairs de ne pas dégager l'excellence avec suffisamment de clarté. C'est la raison invoquée par le RAE pour motiver le passage de l'évaluation par les pairs au tout quantitatif. Ce dernier argument est difficilement recevable : s'il y a une sphère qui, dans chaque domaine, est facilement identifiable à un moment donné, c'est bien celle des contributions majeures. Certes les nobélisables vont généralement avoir des indices de citation élevés, mais on les reconnaîtra en soulignant leurs contributions à la science, pas leurs indices. Il semble donc que l'évaluation quantitative serve avant tout de support dans le pilotage de la recherche menée par le gros des chercheurs en indiquant les travaux et les orientations qui seront soutenus. C'est précisément là que réside le danger.

En effet, la substitution des indices à une évaluation complète prenant en compte les différentes facettes de l'activité des chercheurs introduit des biais importants. En premier lieu, les bases bibliométriques sont nécessai-

1. Rapport « Bibliometrics and the research excellence framework », *Higher Education Funding Council for England,* Londres, 2008.

rement incomplètes[1] (dans le cas des mathématiques, à peine la moitié des revues y est répertoriée) ; évidemment, peu de chercheurs se risqueront à soumettre leurs travaux à des revues non répertoriées. Ensuite, les pratiques de publication peuvent être très différentes, même dans des sous-domaines très peu éloignés – par exemple, il peut être d'usage de publier plusieurs notes par an dans un domaine alors que dans un autre domaine pourtant proche on ne publie que des articles substantiels de fond, donc beaucoup moins nombreux, ce qui affecte considérablement les indices. Même au sein d'un même domaine, certaines orientations peuvent être temporairement plus populaires que d'autres, sans que cela ne préjuge de celle qui sera la plus fructueuse ; un chercheur aura cependant intérêt à rejoindre l'orientation dominante pour maximiser ses indices. Charcuter la publication des résultats ou, plus largement, l'existence de stratégies de citations, comme la constitution de réseaux où l'on s'emploie à se citer et à pousser mutuellement ses articles dans des journaux à haut IF, devient de plus en plus fréquent. Parallèlement, plusieurs études indiquent que la majorité des références citées dans un article ne reflète pas l'influence qu'elles ont eue sur les résultats présentés mais sont d'ordre rhétorique[2], ce qui a tendance à augmenter le taux de citation d'articles déjà très cités.

Le recours exclusif aux indices quantitatifs pour définir la « bonne » recherche, celle qui doit être soutenue et promue, donnera peut-être l'illusion aux tutelles politiques d'un haut rendement par euro investi, mais il risque de se révéler catastrophique pour le progrès des connaissances sur le long terme. Une large partie de la communauté scientifique est consciente de ces dangers. Ainsi, plusieurs sociétés savantes de mathématiciens ont publié un rapport conjoint[3] démontant la pertinence des facteurs de citation lorsque ceux-ci sont utilisés de manière simpliste à la place d'une expertise conduite par des personnes scientifiquement compétentes. Le

1. À noter qu'en pratique il est impossible pour un évaluateur d'établir exactement les indices de citation d'un chercheur : les homonymies (d'autant plus nombreuses que seules les initiales du prénom apparaissent dans les bases), les changements d'affectation, les erreurs de référencement requièrent un tri et des calculs manuels qui rendent le résultat approximatif.

2. C'est-à-dire des références qui ne sont pas en lien direct avec les résultats présentés mais qui sont attendues (livre ou article prestigieux), qui servent à placer son travail (alors que sur le fond il est éloigné de la référence citée), à se signaler auprès de collègues en vue, etc.

3. Rapport « Citation Statistics – A report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS) », juin 2008, disponible sur le site Internet www.mathunion.org/publications.

Plan stratégique du CNRS[1] contient une ferme mise en garde : *« Les dérives visant à donner à la bibliométrie un rôle prépondérant, voire exclusif, s'accompagneraient d'un certain formatage des carrières et d'effets pervers pour l'activité de recherche : minimisation de la prise de risque scientifique, minimisation de la mobilité thématique, frein aux échanges public-privé, stratégies de citations. »* Encore faut-il que les pratiques dans les instances d'évaluation, ainsi que l'organisation de ces instances par les tutelles politiques et les directions des organismes permettent d'échapper à l'emprise des indices quantitatifs et à la manie des classements. Les publications sont avant tout un moyen pour les scientifiques d'exposer leurs résultats et de discuter leurs idées. Une évaluation de qualité ne peut se faire sans comprendre les enjeux scientifiques propres aux travaux examinés, ni sans tenir compte de leur spécificité.

1. Plan stratégique du CNRS « Horizon 2020 », adopté par le conseil d'administration du CNRS le 1er juillet 2008, p. 53. Les différents conseils scientifiques et instances du CoNRS ont également fait part de leur préoccupation dans plusieurs recommandations et motions publiées en 2007 et 2008.