Le manifeste de Leiden pour la mesure de la recherche

The Leiden manifesto for research Metrics, Nature, 520, 429-431

by Diana Hicks, Paul Wouters, Ludo Waltman, Sarah de Rijcke & Ismael Rafols.

http://www.leidenmanifesto.org/

Diana Hicks *est professeur de politique publique au Georgia Institute of Technology d'Atlanta (Géorgie) aux États-Unis.* **Paul Wouters** *est professeur de scientométrie et directeur,* **Ludo Waltman** *est chercheur et* **Sarah de Rijcke** *est 'assistant professor' au Centre for Science and Technology Studies de l'université de Leiden, aux Pays-Bas.* **Ismael Rafols** *est chercheur en politique scientifique au Conseil supérieur de la recherche scientifique, à l'Université polytechnique de Valence, en Espagne et à l'Observatoire des Sciences et Techniques du Haut conseil de l'évaluation de la recherche et de l'enseignement supérieur, à Paris.*

Version française par Ghislaine Filliatreau, directrice de l'Observatoire des Sciences et Techniques du Haut Conseil à l'Evaluation de la Recherche et de l'Enseignement supérieur.
http://www.obs-ost.fr/manifesto

Les chiffres sont de plus en plus utilisés pour piloter la recherche. Les évaluations de la recherche qui étaient auparavant faites « à façon » par des pairs sont aujourd'hui routinisées et s'appuient sur des mesures[1]. Le problème est qu'à présent ces évaluations ne sont plus fondées sur des réflexions mais sur des indicateurs. Et les indicateurs utilisés se sont multipliés : des indicateurs généralement bien intentionnés, pas toujours bien renseignés et souvent mal appliqués. Nous risquons de nuire au système en utilisant les outils mêmes qui ont été conçus pour l'améliorer, car les évaluations sont de plus en plus souvent réalisées par des structures qui n'en ont aucune connaissance ou ne bénéficient d'aucun conseil en matière de bonnes pratiques et d'interprétation.

Avant 2000, on s'appuyait sur le *Science Citation Index* via un CD-ROM de *l'Institute for Scientific Information* (ISI), utilisé par les experts pour des analyses spécialisées. En 2002, Thomson Reuters a lancé une plateforme en ligne, qui rend la base de données Web of Science accessible au plus grand nombre. Depuis lors, deux autres sources de citation sont apparues : *Scopus* de Elsevier (lancé en 2004), et *Google Scholar* (version bêta lancée en 2004). Des outils en ligne permettant de comparer facilement la productivité et l'impact de la recherche institutionnelle ont été introduits, comme InCites (utilisant le Web of Science) et SciVal (utilisant Scopus), ainsi qu'un logiciel permettant d'analyser les profils de citation des auteurs à l'aide de Google Scholar (Publish or Perish, lancé en 2007).

En 2005, Jorge Hirsch, un physicien de l'université de San Diego en Californie a proposé le « h-index » popularisant ainsi le comptage de citations pour les chercheurs. De même, l'intérêt pour le facteur d'impact des revues scientifiques n'a cessé d'augmenter depuis 1995 (voir encart « L'obsession pour le facteur d'impact »).

Plus récemment, les mesures liées à l'usage social et aux commentaires en ligne ont pris de l'ampleur : F1000Prime a été créé en 2002, Mendeley en 2008 et Altmetric.com (soutenu par Macmillan Science and Education, qui détient le groupe Nature Publishing) en 2011.

En tant que scientomètres, spécialistes des sciences sociales et administrateurs de la recherche, nous avons observé avec une inquiétude grandissante le mauvais usage des indicateurs dans l'évaluation de la performance scientifique. Les exemples qui suivent ne sont qu'une fraction de ce que l'on peut observer. Les universités du monde entier sont devenues obsédées par leur position dans les

classements académiques internationaux (comme le classement de Shanghai et la liste du *Times Higher Education*), même si celles-ci sont basées, selon nous, sur des données inexactes et des indicateurs arbitraires. Certains recruteurs exigent que les candidats fournissent leur *h*-index.

Des universités basent leurs décisions de promotion sur des valeurs-seuils de *h*-index et sur le nombre d'articles publiés dans des revues scientifiques « à fort impact ». Les CV des chercheurs sont devenus l'occasion de se vanter de ces scores, particulièrement en biomédecine. Partout, les directeurs de thèse demandent à leurs doctorants de publier dans des revues « à fort impact » et d'obtenir des financements externes avant même d'y être prêts.

Dans les pays scandinaves et en Chine, certaines universités allouent des subventions de recherche ou des primes sur la base d'un indicateur: par exemple, en utilisant le score d'impact individuel pour affecter des « financements à la performance », ou en accordant une prime aux chercheurs qui publient dans une revue ayant un facteur d'impact supérieur à 15[2].

Dans de nombreux cas, les chercheurs et les évaluateurs continuent à faire preuve de discernement. Pour autant, l'abus d'indicateurs concernant la recherche est devenu trop répandu pour être ignoré.

Nous présentons donc le Manifeste de Leiden, du nom de la conférence au cours de laquelle il a vu le jour (voir http://sti2014.cwts.nl). Ses dix principes ne sont pas nouveaux pour les scientomètres, même si aucun d'entre nous n'avait jusqu'ici pu les énoncer aussi clairement, faute de langage commun. D'éminents chercheurs dans ce domaine, comme Eugene Garfield (fondateur de l'ISI), ont déjà énoncé certains de ces principes[3,4]. Mais ces experts ne sont pas là quand les évaluateurs rédigent leur rapport d'évaluation à l'attention d'administrateurs qui, eux-mêmes, ne sont pas des experts de la méthodologie. Et les chercheurs qui cherchent des textes de référence pour contester une évaluation ne trouvent que des informations éparses dans ce qui est, pour eux, des revues aux contenus complexes auxquelles ils n'ont pas forcément accès.

Aussi nous proposons ici un condensé des bonnes pratiques en matière d'évaluation de la recherche basée sur les indicateurs, afin que les chercheurs puissent demander des explications aux évaluateurs, et que les évaluateurs puissent interroger l'exactitude de leurs indicateurs.

LES DIX PRINCIPES

1. La description quantitative doit servir d'appui à une évaluation qualitative par des experts. Les indicateurs quantitatifs permettent de contrebalancer les biais liés aux évaluations par les pairs, et ils facilitent les délibérations. Ce qui devrait améliorer l'évaluation par les pairs, dans la mesure où il est difficile de porter un jugement sur les travaux de ses collègues sans disposer d'un certain nombre d'informations pertinentes. Cependant, les évaluateurs ne doivent pas céder à la tentation de laisser de simples chiffres dicter leurs décisions. Les indicateurs ne doivent pas se substituer à des réflexions éclairées. Chacun reste responsable de ses jugements.

2. Mesurer la performance au regard des missions de recherche de l'institution, du groupe ou du chercheur. Les objectifs assignés à un programme de recherche doivent être énoncés dès le départ, et les indicateurs utilisés pour évaluer leur performance doivent être clairement en lien avec ces objectifs. Le choix des indicateurs, et la manière dont ils sont utilisés, devraient tenir compte du contexte socio-économique et culturel. Les scientifiques peuvent avoir des missions de recherche diverses et variées. La recherche qui repousse les limites des connaissances scientifiques diffère de la recherche qui vise à fournir des solutions à des problèmes sociétaux. L'évaluation peut être davantage basée sur la pertinence des recherches pour les politiques publiques, l'industrie ou la société que sur une notion d'excellence académique. Il n'existe aucun modèle d'évaluation qui puisse s'appliquer à tous les contextes.

3. Protéger l'excellence dans les domaines de recherche importants à l'échelle locale. Dans de nombreuses parties du monde, l'excellence en matière de recherche sous-entend des publications en anglais. La loi espagnole, par exemple, fait explicitement mention du souhait que les chercheurs espagnols publient dans des revues à fort impact. Le facteur d'impact est calculé pour les revues indexées dans la base de données Web of Science, maintenue depuis les États-Unis et toujours principalement en anglais. Ces biais sont particulièrement problématiques pour les sciences sociales et humaines, pour lesquelles les recherches sont souvent menées à l'échelle régionale ou nationale. De nombreux autres domaines de recherche ont une dimension nationale ou régionale, par exemple, l'épidémiologie du VIH en Afrique subsaharienne.

Ce pluralisme et cette importance sociétale tendent à être minorés en faveur de la rédaction d'articles n'intéressant que les « gardiennes des publications à fort impact » : les revues anglophones. Les sociologues espagnols les plus cités dans le Web of Science ont travaillé sur des modèles abstraits ou étudié des données américaines. Sont ainsi perdus de vue des travaux spécifiques menés par des sociologues publiant dans des revues à fort impact hispanophone, sur des sujets tels que le droit du travail local, les soins de santé en milieu familial ou l'emploi des immigrés[5]. Les indicateurs fondés sur des revues de haute qualité non anglophones devraient servir à identifier et à récompenser l'excellence dans des domaines de recherche d'intérêt local.

4. Maintenir une collecte de données et des processus d'analyse ouverts, transparents et simples. La construction des bases de données nécessaires aux évaluations doit suivre des règles claires, formulées avant que le travail ne soit terminé. C'était la pratique commune en recherche et dans les structures privées qui ont mis au point, sur plusieurs décennies, les méthodologies d'évaluation bibliométrique. Ces groupes ont publié leurs protocoles dans des revues examinées par des collègues, une transparence qui permettait un examen approfondi. En 2010 par exemple, un débat public sur les propriétés techniques d'un indicateur important utilisé par l'un des groupes du Manifeste (le Centre for Science and Technology Studies de l'université de Leiden, aux Pays-Bas) a conduit à une révision du calcul de cet indicateur[6]. Les nouveaux entrants issus du secteur privé devraient suivre les mêmes standards ; personne ne devrait accepter d'être évalué par une boîte noire.

La simplicité est une vertu pour un indicateur, parce qu'elle renforce la notion de transparence. Mais des indicateurs simplistes peuvent déformer les choses (voir le principe 7). Les évaluateurs doivent s'efforcer de trouver un équilibre en s'appuyant sur des indicateurs simples reflétant fidèlement la complexité du processus de recherche.

5. Permettre aux chercheurs évalués de vérifier les données et analyses. Afin de garantir la qualité des données, tous les chercheurs concernés par des études bibliométriques devraient pouvoir vérifier que leurs productions ont été correctement identifiées. Toute personne responsable de la direction et de la gestion des processus d'évaluation devrait pouvoir garantir l'exactitude des données, par le biais d'une auto-vérification ou d'un audit externe. Les universités devraient pouvoir appliquer ces règles à leurs systèmes d'information, et elles devraient constituer pour eux un principe directeur dans le choix des fournisseurs de ces systèmes. La collecte et le traitement de données précises et de haute qualité requièrent du temps et de l'argent. Ils doivent être budgétés.

6. Tenir compte des différences entre disciplines en matière de publication et de citation. Une bonne pratique consiste à sélectionner un ensemble d'indicateurs possibles pour permettre aux différentes disciplines de recherche de choisir les plus adaptés à leur champ de recherche. Il y a quelques années de cela, un groupe européen d'historiens avait reçu une note relativement faible dans le cadre d'une évaluation nationale par leurs pairs, car ils avaient écrit des livres et non des articles indexés par le Web of Science. Ces historiens avaient la malchance de faire partie d'un

département de psychologie. Les historiens et les chercheurs en sciences sociales souhaitent que les livres et les articles publiés dans les revues nationales soient pris en compte dans le décompte de leurs publications ; de même, les chercheurs en sciences informatiques souhaitent que les textes de leurs conférences soient pris en compte.

Les taux de citations varient selon les domaines : les revues mathématiques les mieux classées ont un facteur d'impact d'environ 3 ; les revues de biologie cellulaire les mieux classées ont un facteur d'impact d'environ 30. Il est donc nécessaire d'utiliser des indicateurs normalisés. La méthode de normalisation la plus fiable est basée sur les percentiles : chaque article est pondéré sur la base du percentile dans lequel il se trouve dans la répartition des citations de ce domaine (les top 1 %, 10 % ou 20 %, par exemple). Une seule publication à fort taux de citation améliore légèrement la position d'une université dans un classement basé sur des indicateurs en percentile, mais dans le cas d'un classement basé sur des moyennes de citations, la même publication peut faire passer l'institution du milieu au haut du classement[7].

7. Baser les évaluations des chercheurs sur un jugement qualitatif de leurs travaux. Plus vous êtes âgé, plus votre indice *h* est élevé, même si vous ne publiez pas de nouveaux articles. Le h-index varie selon les domaines : celui des chercheurs en science de la vie atteint 200, celui des physiciens 10 et celui des chercheurs en sciences sociales 20-30[8]. Il dépend également des bases de données : certains chercheurs en sciences informatiques se voient attribués un indice *h* d'environ 10 dans le Web of Science, et de 20-30 dans Google Scholar[9]. Lire et juger les travaux d'un chercheur est toujours plus pertinent que se baser uniquement sur un chiffre. Même lorsque l'on compare un grand nombre de chercheurs, il est toujours préférable d'adopter une approche tenant compte d'informations complémentaires relatives à l'expertise, à l'expérience, aux activités et à l'influence de chaque individu.

8. Éviter les simplifications abusives et les fausses précisions. Les indicateurs en science et technologie présentent une certaine tendance à l'ambiguïté conceptuelle et à l'incertitude, et nécessitent des hypothèses fortes qui ne sont pas forcément acceptées par l'ensemble de la communauté. La signification des décomptes de citations, par exemple, est débattue depuis longtemps. C'est pourquoi une bonne pratique consiste à utiliser plusieurs indicateurs afin d'aboutir à une description plus précise et transversale. Dans les cas où les incertitudes et erreurs peuvent être quantifiées, par exemple à l'aide de barres d'erreur, les valeurs d'indicateurs publiées devraient être accompagnées de ce type d'informations. Dans le cas contraire, les personnes calculant les indicateurs devraient au moins veiller à éviter les fausses précisions. Par exemple, le facteur d'impact d'une revue est publié à trois décimales près pour éviter les ex-æquo. Cependant, étant donné l'ambiguïté conceptuelle et la variabilité aléatoire des décomptes de citations, cela n'a aucun sens de distinguer deux revues sur la base de leur très faible différence de facteur d'impact. Il faut éviter toute fausse précision : une précision à plus d'une décimale près n'est pas justifiée.

9. Reconnaître les impacts systémiques des évaluations et des indicateurs. Les indicateurs impactent le système par les pratiques qu'ils encouragent. Ces impacts doivent être anticipés. Cela signifie qu'il est toujours préférable d'utiliser un ensemble d'indicateurs : un seul indicateur risque de conduire à des comportements de triche ou d'altération de l'objectif poursuivi (la mesure devenant le nouvel objectif). Dans les années 1990 par exemple, l'Australie a subventionné la recherche universitaire en utilisant une formule basée en grande partie sur le nombre d'articles publiés par les institutions. Les universités pouvaient calculer la « valeur » d'un article dans une revue référencée ; en 2000, cette subvention était fixée à 800 dollars australiens (480 dollars US). Comme on pouvait s'y attendre, le nombre d'articles publiés par les chercheurs australiens a augmenté, mais il s'agissait de revues moins citées, suggérant une baisse de la qualité de ces articles[10].

10. Réévaluer régulièrement et faire évoluer les indicateurs. Les missions de la recherche et les objectifs des évaluations évoluent, entraînant la transformation du système de recherche lui-même. Des mesures autrefois utiles deviennent obsolètes, et de nouvelles les remplacent. Les systèmes d'indicateurs doivent être révisés et parfois modifiés. S'étant rendu compte des effets de sa formule trop simpliste, l'Australie a adopté, en 2010, un programme plus complexe, *Excellence in Research for Australia*, qui met l'accent sur la qualité.

PROCHAINES ÉTAPES

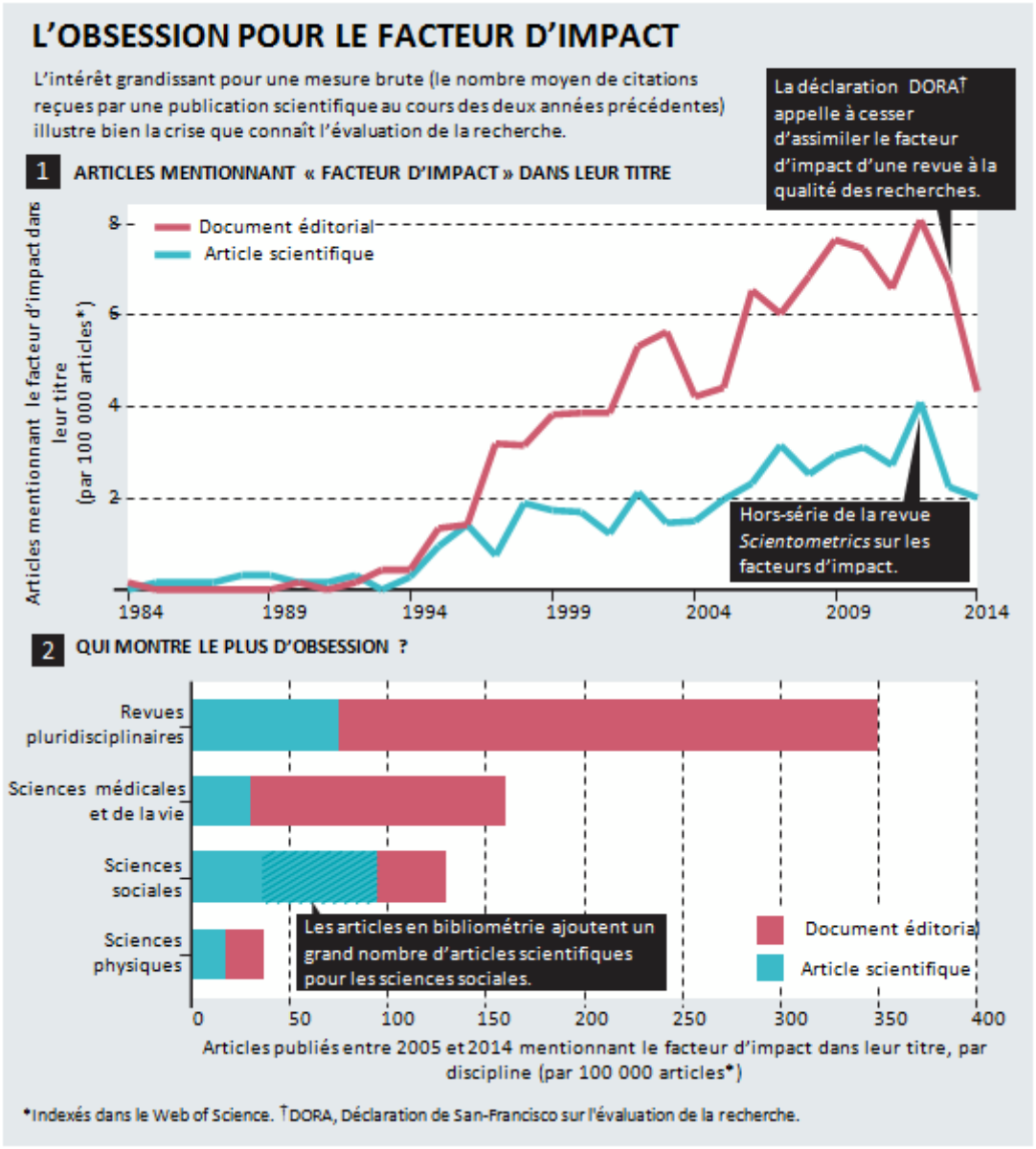
Guidée par ces dix principes, l'évaluation de la recherche devrait pouvoir jouer un rôle important dans le développement de la science et de ses interactions avec la société. Les indicateurs sur la recherche peuvent fournir des informations essentielles qu'il serait difficile de collecter ou de comprendre par des expertises individuelles. Mais il faut veiller à ne pas faire de l'instrument que constituent ces informations quantitatives un objectif, une fin en soi.

Les meilleures décisions sont prises en associant des méthodes de calcul fiables et une capacité à apprécier la portée et la nature des recherches évaluées. Des éléments d'appui à la fois quantitatifs et qualitatifs sont nécessaires ; chacun est objectif à sa façon. La décision en matière de science doit être fondée sur des processus rigoureux, renseignés par des données de qualité.

Références :

1. Wouters, P. dans *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact* (Éds Cronin, B. & Sugimoto, C.) 47–66 (MIT Press, 2014).
2. Shao, J. & Shen, H. *Learned Publishing* **24**, 95–97 (2011).
3. Seglen, P. O. *Br. Med. J.* **314**, 498–502 (1997).
4. Garfield, E. *J. Am. Med. Assoc.* **295**, 90–93 (2006).
5. López Piñeiro, C. & hicks, D. *Res. Eval.* **24**, 78–89 (2015).
6. van Raan, A. F. J., van Leeuwen, T. N., Visser, M. S., van Eck, N. J. & Waltman, L. *J. Informetrics* **4**, 431–435 (2010).
7. Waltman, L. *et al. J. Am. Soc. Inf. Sci. Technol.* **63**, 2419–2432 (2012).
8. Hirsch, J. E. *Proc. Natl Acad. Sci. USA* **102**, 16569–16572 (2005).
9. Bar-Ilan, J. *Scientometrics* **74**, 257–271 (2007).
10. Butler, L. *Res. Policy* **32**, 143–155 (2003).

Illustration :



SOURCE DES DONNÉES : THOMSON REUTERS, WEB OF SCIENCE ; ANALYSE : D.H., L.W.