CONTEXTE Le 22 janvier 2009, le président de la République a dressé un portrait peu flatteur de la production scientifique française. Son discours très critique sur la place de la France dans les palmarès mondiaux de la recherche a déclenché de vives réactions dans le monde universitaire et scientifique. En cause, la prédominance d'un classement mondial des universités réalisé depuis quelques années par des chercheurs de l'université Jiao-Tong de Shanghai. Une partie des réformes de la recherche et de l'enseignement supérieur en cours semblent avoir pour but de rehausser la place des universités françaises dans ce classement.

Le classement de Shanghai n'est pas scientifique

La place de la France dans les palmarès mondiaux influence les réformes en cours. Ainsi « évaluer » est devenu le mot à la mode dans la recherche et l'enseignement supérieur français. Encore faut-il utiliser des indicateurs fiables et pertinents.

Yves Gingras
est professeur au département d'histoire de l'université du Québec à Montréal (UQAM), titulaire de la chaire de recherche du Canada en histoire et sociologie des sciences et directeur scientifique de l'Observatoire des sciences et des technologies de l'UQAM.

Depuis quelques années, le monde académique européen semble atteint d'une véritable fièvre d'évaluation de la recherche et des universités. En France, notamment, 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 à ces questions. 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 indicateurs de la production scientifique. Ajoutées aux habituels (et donc mieux connus) investissements en recherche et développement qui sont des mesures d'*input* de la recherche, les publications servent de mesure d'*ouput* – et les citations qu'elles reçoivent constituent un indice de leur visibilité internationale et, indirectement, de leur « qualité » et de leur « impact » scientifique.

Comme le montre très bien le rapport du sénateur Joël Bourdin, rendu public en juillet 2008, les différents classements ont chacun leurs limites et manifestent la fâcheuse tendance à valoriser systématiquement les universités de certains pays [1]. Ainsi le classement dit de Shanghai est très favorable aux universités américaines ; le classement anglais favorise, quant à lui, les performances des établissements du Royaume-Uni ; et le classement de Leiden donne de belles places aux universités néerlandaises. L'auteur aurait pu ajouter que le classement de l'école des Mines favorise les grandes écoles françaises [2]. En d'autres termes, il est toujours possible de trouver un indicateur qui nous avantage. 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 il faut les remplacer par d'autres, plus adéquats. Utiliser des classements fondés sur de mauvaises mesures a des effets pervers en stimulant des politiques qui s'appuient sur des problèmes mal identifiés.

Manie des classements

Ainsi il est pour le moins curieux d'apprendre que 61 % des dirigeants d'établissements de l'enseignement supérieur français disent vouloir améliorer leur rang dans le fameux classement de Shanghai alors qu'ils ignorent probablement ce qu'il mesure vrai-

ment ! 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* [3] » sans s'assurer que ce classement est bien valide. Or, on le verra plus loin, il ne possède en fait aucune des propriétés que doit posséder un bon indicateur.

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 locales [4]. Or, certains critiquent avec raison un tel classement, fondé sur un panel d'experts choisis on ne sait trop comment, 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, unidimensionnels, et plus ou moins occultes quant à leur méthode [5]. Ce classement, parfois confondu avec la bibliométrie, montre la confusion qui existe dans les esprits des uns et des autres entre « évaluation » et « 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 certains tendent à « jeter le bébé avec l'eau du bain » : ils confondent les usages simplistes de cet outil avec l'outil lui-même.

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 en utilisant comme indicateurs les publications scientifiques et les citations qu'elles contiennent.

Longtemps, seule la société Thomson Reuters proposait des bases de données bibliographiques. Ce monopole historique explique que la plupart des travaux de scientométrie reposent sur les bases de données de cette société. Depuis 2004, une nouvelle banque de données intitulée Scopus, mise sur le marché par Elsevier, couvre davantage de revues (environ 16 000, toutes disciplines confondues) et fait directement concurrence à Thomson [6]. L'intérêt de ces sources concernant l'évaluation est qu'elles sont contrôlées et que l'on connaît la liste des revues qui y sont recensées. Le problème, bien sûr, est qu'elles ne sont pas gratuites…

© RAPHAËL FOURNIER/FEDEPHOTO

EN 2009, LES MOBILISATIONS des enseignants-chercheurs et des chercheurs du CNRS se sont en partie cristallisées sur la question de l'évaluation tout au long de la carrière professionnelle.

Anarchie évaluative

Google Scholar, et même Internet sont gratuits, et servent de plus en plus comme bases de données pour l'analyse bibliométrique (et par extension « webométrique »). Mais ces deux sources sont non contrôlées et non reproductibles : 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.

L'accès gratuit à Google Scholar et à Internet a contribué à ce climat d'anarchie évaluative. Tout chercheur est en effet tenté d'y mesurer sa « visibilité » ou sa « qualité ». D'où une multiplication des usages spontanés de l'évaluation et créations de prétendus palmarès et indicateurs de l'impact de la recherche qui contribuent à créer un certain chaos dans le monde académique.

En fait, l'existence et la persistance de ces indicateurs et palmarès divers et variés relèvent de l'adage selon lequel « *any number beats no number* » (mieux vaut ⇨

[1] www.senat.fr/rap/r07-442/r074421.pdf

[2] www.ensmp.fr/PR/defclassementEMP.pdf

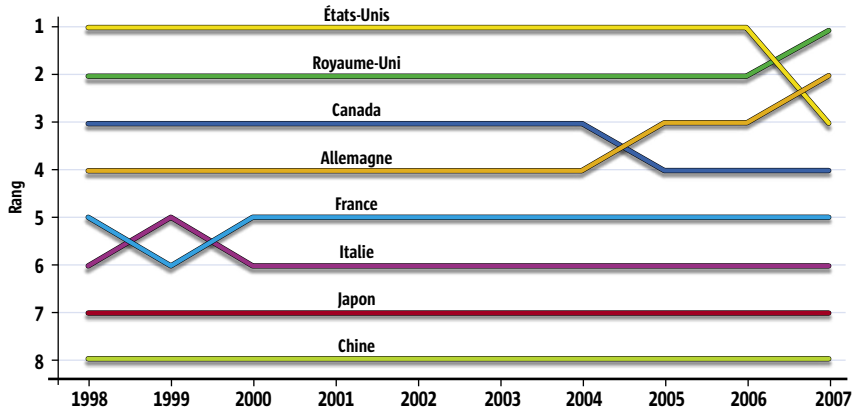
[3] Les cahiers de la compétitivité. Spécial enseignement supérieur, p. II ; encart dans *Le Monde* du 21 mai 2008.

[4] http://tinyurl.com/ddrmlu

[5] www.sauvonsluniversite.com/spip.php?article591

[6] E. Archambault *et al.*, *Journal of the American Society for Information Science and Technology*, sous presse.

Fig.1 Le G8 des citations



SELON CET INDICATEUR fondé sur le nombre relatif de citations des articles publiés dans chacun des huit pays mentionnés, la France est en position stable dans toutes les disciplines depuis dix ans. On ne peut donc pas parler de déclin de l'impact de la recherche française. Bien que la Chine ait fortement augmenté sa production brute d'articles, sa position n'a pas bougé en termes d'impact.

SOURCE SCIENCE CITATION INDEX, THOMSON-REUTERS. DONNÉES COMPILÉES PAR OST-UQAM

⇨ n'importe quel chiffre que pas de chiffre du tout !) Deux exemples emblématiques : le classement de Shanghai et le « *h* index ». Le premier interpelle et fascine les décideurs politiques et les présidents d'université car il propose un classement mondial des universités. Le second circule plutôt parmi les scientifiques eux-mêmes et « évalue » les chercheurs au niveau individuel. Dans les deux cas, on utilise un seul nombre (obtenu en combinant différents calculs) pour classer et évaluer la qualité de la recherche, et ce, malgré le caractère multidimensionnel de la recherche.

Mais au fait, qu'est-ce qu'un indicateur bien construit ? Un indicateur est – rappelons-le – une variable qui vise à appréhender un concept. L'indicateur n'est pas le concept lui-même mais une façon approchée de mesurer les modifications temporelles dudit concept. La première propriété d'un bon indicateur est donc 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 ?

Le niveau d'investissement en R & D d'un pays est, par exemple, une première (bonne) mesure de l'intensité de la recherche dudit pays. Mais supposons que l'on veuille mesurer l'impact scientifique d'un auteur. On peut penser que le nombre total de citations obtenues par un chercheur peut servir d'indicateur. Cependant, 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 de la qualité 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 des citations obtenues par un chercheur et la renommée mesurée par des prix obtenus ou des nominations académiques [7]. Notons toutefois que cet indice des citations a été surtout validé en sciences de la nature. On ne peut le transférer sans précautions dans les sciences sociales – et encore moins dans les lettres –, car ces disciplines utilisent davantage le livre que l'article comme mode de diffusion des résultats [8].

Indicateur homogène

Seconde propriété d'importance, un indicateur se doit d'être homogène. Exemple : un indicateur homogène (à l'échelle d'un pays) 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* qui peut aussi être comparée à une mesure d'*input*, comme la valeur des investissements en recherche. Ces indicateurs permettent de comparer les pays, et même les institutions entre elles. Ils peuvent aussi servir à construire une cartographie des activités selon deux mesures différentes : les *inputs* et les *outputs*. Le rapport de ces deux mesures fournit un indice composite de productivité (*input/output*). En revanche, si l'on prend cet indicateur comme une mesure de « qualité » de la recherche (et non de son efficience) ou qu'on le combine avec un indicateur de réputation fondé sur un panel d'experts, alors on obtient un indicateur assez

PUBLICATIONS

La Chine devant la France ?

LA PROPORTION DU PIB d'un pays consacré à la recherche et au développement (R & D), publiée depuis longtemps par l'OCDE offre une 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 produisent essentiellement le même classement, au moins pour les 25 pays les plus importants [1]. De plus, il existe une corrélation très forte entre le nombre d'articles publiés par pays et son niveau de dépenses en R & D. En termes de nombre total de publications produites en 2005 par les 8 pays les plus productifs (figure ci-contre, hors États-Unis, au premier rang), la France se classe au 6e 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 trois ou cinq 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

hétérogène qui pourra varier de façon imprévue, et qui n'aura pas de signification claire.

Enfin, un indicateur de qualité varie en conformité avec l'inertie propre de l'objet mesuré. Qu'entend-on par là ? Plaçons un thermomètre dans une pièce fermée et supposons qu'au lieu de la bonne vieille colonne de mercure ou d'alcool, on utilise un instrument électronique à écran numérique. Ce thermomètre indique 20 degrés, puis une minute plus tard 12, et encore une minute plus tard 30. Il est certain que le bon sens force l'observateur à conclure que l'instrument est défectueux, car 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 frivoles). En conséquence, un palmarès annuel qui montrerait qu'une institution est passée en une seule année du 1er au 6e rang ou du 8e au 2e rang suggérerait fortement que l'indicateur utilisé est défectueux ou trop imprécis, et non pas que la qualité de l'institution ait chuté ou augmenté soudainement ! 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 si rarement ? Parce qu'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.

Jetons maintenant un coup d'œil sur le palmarès de Shanghai et sur l'indice *h* (aussi appelé « *h* index »). Le premier évalue des institutions tandis que le second évalue des individus. Attendu chaque année avec impatience par de nombreux dirigeants d'université, le classement de Shanghai des supposées « meilleures » universités mondiales est composé de six mesures (quatre ont un poids de 20 % et deux de 10 %) : le nombre de membres du corps universitaire ayant reçu un Nobel ou une médaille Fields (pour les mathématiques) ; le nombre de chercheurs de l'institution parmi la liste des « plus cités » de Thomson Reuters ; le nombre d'articles de l'institution publiés dans les revues *Nature* et *Science ;* le nombre total d'articles recensés dans le Web of Science de la compagnie Thomson Reuters ; le nombre d'anciens étudiants ayant reçu un Nobel ou une médaille Fields ; l'ajustement des résultats précédents selon la taille de l'institution.

On voit donc que l'indice final attribué à une institution se fonde sur la somme de plusieurs mesures hétérogènes, car le nombre de publications dans les revues *Science* et *Nature* n'est pas commensurable au nombre de prix Nobel. Chose plus surprenante, il a été montré que les classements obtenus sont difficilement reproductibles, lorsque l'on retourne aux données disponibles sur les bases de données [9]. Par ailleurs, on peut aussi mettre en question le choix d'une mesure 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, 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 [10] ! Sans parler du fait que l'on ⇨

L'indice final attribué à une institution se fonde sur des mesures hétérogènes

[7] J.R Cole et S. Cole, *Social Stratification of Science,* Chicago University Press, 1973.

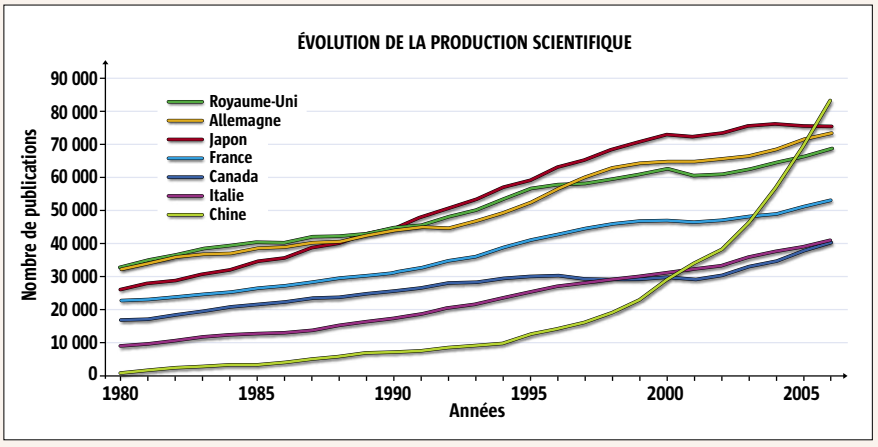
[8] V. Larivière *et al., Journal of the American Society for Information Science and Technology, 59,* 288, 2008.

[9] R.V. Florian, *Scientometrics, 72,* 25, 2007.

[10] M. Enserink, *Science, 317,* 1026, 2007.

5e rang, toujours pour 2005. La différence s'explique aisément : pour le nombre de publications, la Chine passe avant la France, mais pas pour ce qui concerne les citations reçues par ses articles. De même, le Canada, au 8e rang pour la production, se classe au 4e rang en termes de citations.

[1] E. Archambault *et al., Book of Abstracts of the 10th International Conference on Science and Technology Indicators,* 2008.



[11] J.E. Hirsch, *PNAS*, *102*,16569, 2005.

[12] G. Filiatreau (dir.), *Indicateurs des sciences et des technologies*, Economica, 2008.

⇨ peut se demander si la qualité d'une université en 2006 peut être influencée par des travaux effectués plus de quatre-vingts ans auparavant…

Bref, ce classement si couru n'a, en réalité, aucune valeur scientifique. Il est probable que l'importance accordée à ce classement soit un effet des discours sur l'internationalisation du « marché universitaire » et de 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 » qu'il ne faudrait pas laisser aux seules universités américaines. Enfin, ceux qui veulent réformer le système universitaire se servent du palmarès de Shanghai de façon opportuniste pour justifier leurs politiques. Parions que, si les universités françaises étaient très bien classées, il aurait été plus difficile de justifier les réformes en cours. Parions aussi que les décideurs auraient alors jeté un regard plus critique sur le palmarès…

Passons maintenant à l'indice *h*. Construit par le physicien Jorge E. Hirsch, de l'université de Californie à San Diego [11], l'indice *h* d'un chercheur est le nombre d'articles *n* qu'il a publiés et qui ont reçu au moins *n* citations (sur une période donnée). Par exemple, un auteur qui a publié vingt articles parmi lesquels dix ont au moins dix citations chacune aura un indice *h* de 10. Cet indicateur de « qualité » 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*. Un tel mélange devrait déjà nous faire douter de la fiabilité de l'indice *h*. Mais, comme s'il contribuait à satisfaire le narcissisme des scientifiques, son usage s'est généralisé en moins de deux ans, et est même incorporé dans certaines banques de données !

Pour mieux saisir l'inutilité, voire la méprise, que cet indice peut générer, comparons deux cas de figure : un jeune chercheur A a publié seulement trois articles, mais ceux-ci ont été cités soixante fois chacun ; un second chercheur du même âge B est plus prolifique et possède à son actif dix articles, cités onze fois chacun sur la même période. Ce second chercheur aura donc un indice *h* de 10, alors que le premier aura un indice de 3. Peut-on en conclure que B est trois fois « meilleur » que A ? Bien sûr que non…

Collègues et bureaucrates

Curieusement, ce sont les scientifiques eux-mêmes qui succombent aux usages anarchiques de la bibliométrie individuelle et qui, siégeant parfois au sein de différents comités et conseils d'administration d'organes décisionnels de la recherche, suggèrent d'en généraliser l'usage. Tout cela confirme que, dans le champ scientifique, l'ennemi est souvent moins le bureaucrate que le collègue.

Cela dit, il est possible de construire des indicateurs agrégés de la recherche qui donnent une bonne idée de la position relative à l'échelle nationale ou mondiale, des universités et des pays dans les grands champs disciplinaires (lire « La Chine devant la France ? » p. 48).

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 4e position en termes de citations relatives, et, en mathématiques, sa position oscille, entre 2000 et 2005, entre la 2e et la 3e position, ce qui confirme la forte tradition des mathématiques en France.

La National Science Foundation publie le *Science and Engineering Indicators* depuis 1972, lequel comprend des données bibliométriques comparatives très utiles. En France, l'Observatoire des sciences et des techniques (OST), créé en 1990, fait lui aussi paraître tous les deux ans un volume intitulé *Indicateurs des sciences et des technologies,* qui inclut des données bibliométriques [12].

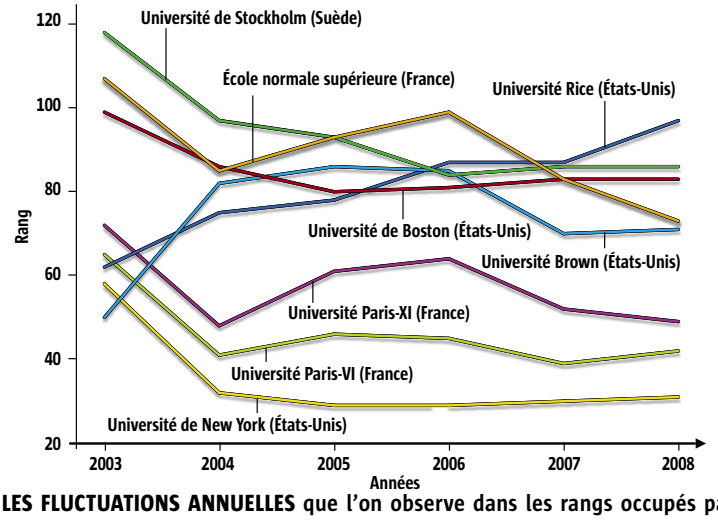
Ces exemples montrent qu'il est possible d'éviter les effets les plus simplistes des classements en organisant de façon adéquate les données. La production et la publication de tableaux bibliométriques comparatifs par domaine de recherche, et même par institution, compilées sur quelques années, peuvent permettre d'analyser sereinement les tendances mondiales ou nationales de la recherche et de prendre des décisions éclairées en fonction des priorités locales ou nationales. En revanche, ce type de données ne peut être utilisé à l'échelle des individus. ■■ Y. G.

POUR EN SAVOIR PLUS

■ **http://tinyurl.com/gingras-eval**

■ Benoît Godin, « James Cattell mesure la science », *La Recherche*, octobre 2006, p. 54.

Fig.2 Un classement trop irrégulier



LES FLUCTUATIONS ANNUELLES que l'on observe dans les rangs occupés par plusieurs universités au classement de Shanghai sont très importantes, parfois 30 places d'une année à l'autre. Elles suffisent à discréditer ce classement : les institutions académiques évoluent dans la réalité beaucoup plus lentement.