



Article professionnel

Article

2011

Published version

Open Access

This is the published version of the publication, made available in accordance with the publisher's policy.

La bibliométrie pour les nuls

Iriarte, Pablo; Krause, Jan Brice

How to cite

IRIARTE, Pablo, KRAUSE, Jan Brice. La bibliométrie pour les nuls. In: Hors-texte, 2011, vol. 95.

This publication URL: <https://archive-ouverte.unige.ch/unige:23117>

LA BIBLIOMÉTRIE POUR LES NULS

Pablo Iriarte, pablo.iriarte@chuv.ch (CHUV/BiUM)
Jan Krause, jan.krause@unige.ch (UNIGE/BFM ; CHUV/BiUM)

Notre métier traverse une période d'évolution fulgurante et les disciplines voisines viennent bousculer nos pratiques. C'est le cas de la bibliométrie, domaine en pleine expansion qui sollicite de plus en plus la participation des professionnels dans les bibliothèques universitaires.

Selon l'article de Wikipedia [1], « la bibliométrie peut être définie comme 'l'application des mathématiques et des méthodes statistiques aux livres, articles et autres moyens de communication ». En d'autres termes, il s'agit d'une analyse quantitative de l'activité des réseaux de communication scientifique. « [Elle] a un volet cognitif, en interaction avec les champs se donnant les sciences et les communautés scientifiques comme objet (économie de la connaissance, sociologie des sciences, épistémologie, histoire des sciences, etc.) et avec les sciences de l'information. Elle a un volet opérationnel, en liaison avec l'évaluation, le positionnement des acteurs et le management scientifique. Dans ce dernier cas, les études bibliométriques permettent de quantifier la communication d'un individu ou d'un groupe, non seulement en termes de volume, mais également de visibilité, d'influence, de partenariats, d'insertion dans les réseaux. » [2]

En Suisse romande comme ailleurs, les institutions académiques s'intéressent au volet opérationnel de cette science. Et cela comporte des aspects délicats : les résultats bibliométriques sont en général utilisés dans l'attribution des financements et plus généralement à l'évaluation des chercheurs et groupes de recherche. Dans ce contexte, des analyses biaisées peuvent avoir des conséquences sévères. Les bibliothécaires, avec leur expertise et leur connaissance des différents acteurs et outils du monde de l'information, ont un rôle à jouer à ce titre. En effet, pour produire une analyse bibliométrique de qualité, il nous semble d'une part crucial de bien connaître les sources de données bibliographiques, leurs lacunes et les sources d'erreur qui leur sont associées. D'autre part, il est nécessaire de comprendre la signification des différentes métriques et calculs qui se cachent derrière les indicateurs. Il faut préciser que le nombre d'indicateurs bibliométriques existants est très grand. Chaque indicateur possède ses avantages, ses inconvénients, et ses biais. Par dessus le marché, ces métriques peuvent être pondérées de diverses manières. Finalement, leur signification varie selon les disciplines scientifiques considérées.

Grosso modo, on distingue deux types d'indicateurs bibliométriques : les indicateurs d'impact « indirect » calculés au niveau des revues et les indicateurs d'impact « direct » se focalisant sur les citations reçues par chaque document publié par un chercheur, un groupe de recherche ou une institution.

[1] Wikipédia : Bibliométrie. <http://fr.wikipedia.org/wiki/Bibliom%C3%A9trie>, consulté le 12.03.2011

[2] Pritchard A., « Statistical Bibliography or Bibliometrics », Journal of Documentation, 25, 4, p. 348-349, 1969

Les indicateurs bibliométriques indirects les plus utilisés sont:

Facteur d'impact (IF) [3]

Indicateur inventé par Eugene Garfield, fondateur de l'Institute for Scientific Information (ISI). L'ISI appartient maintenant à Thomson Reuters. Historiquement, il s'agit du premier grand indicateur bibliométrique et ce depuis 1975. Celui-ci est publié annuellement dans le Journal Citation Reports (JCR) [4], base qui compte actuellement environ 10'000 journaux. L'IF est une mesure de l'impact immédiat d'un journal car il donne la moyenne de citations annuelles reçues par les articles publiés dans les deux années précédentes. Par exemple, l'IF 2010 d'un journal donné est le rapport entre le nombre de citations reçues en 2010 par les articles de ce journal publiés en 2008 et 2009 et le nombre total d'articles publiés dans ce journal pendant ces deux ans.

Research Production Unit (RPU) [5]

Indicateur dérivé de l'IF et dont les valeurs sont normalisées par domaine. Il permet de corriger un biais important de l'IF car les habitudes, les taux et les cinétiques de citations varient d'une discipline à une autre. Par exemple, le nombre de citations moyen dans la bibliographie change d'une discipline à une autre [6].

SCImago Journal Rank (SJR) [7]

Indicateur développé par la compagnie SCImago Lab et dérivé de l'algorithme itératif utilisé par Google pour calculer le PageRank. Couvrant les citations comptabilisées dans la base de données SCOPUS (Elsevier), il s'agit du grand concurrent de l'IF. Son intérêt principal réside dans sa pondération des citations : celles provenant des journaux importants ont une valeur plus grande que celles des journaux plus modestes.

Eigenfactor et Article Influence Score [8]

Ces deux métriques reposent sur les données de citations de Thomson Reuters pour calculer l'impact des journaux. Elles prennent en compte la structure des citations. Les liens entre les journaux constitués par les citations forment une carte, ou graphe, de la production scientifique. Ce graphe est ensuite exploité pour pondérer les

[3] The Thomson Reuters Impact Factor [This essay was originally published in the Current Contents print editions June 20, 1994, when Thomson Reuters was known as The Institute for Scientific Information® (ISI®)]

http://thomsonreuters.com/products_services/science/free/essays/impact_factor

[4] http://thomsonreuters.com/products_services/science/science_products/a-z/journal_citation_reports

[5] Schwartz S. and Lopez Hellin J. Measuring the impact of scientific publications. The case of the biomedical sciences. Scientometrics, 35(1):119-132.

[DOI:10.1007/BF02018236](https://doi.org/10.1007/BF02018236)

[6] Adam D. Citation analysis: The counting house. Nature. [10.1038/415726a]. 2002;415(6873):726-9.

[7] SCImago Journal & Country Rank: Journal Rankings.

<http://www.scimagojr.com/journalrank.php>

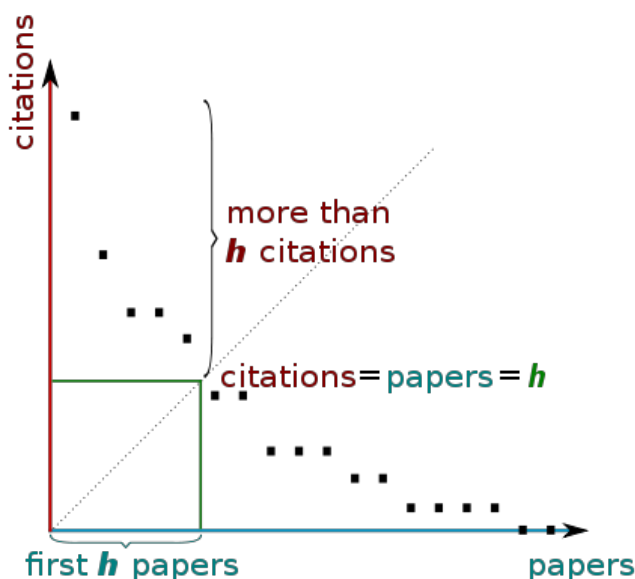
[8] <http://www.eigenfactor.org>

résultats. Ces deux métriques se déploient sur une fenêtre de 5 ans et sont considérées comme plus stables que l'IF (calculé sur 2 ans).

Les indicateurs indirects sont encore très utilisés pour l'évaluation de la production scientifique aux niveaux institutionnels et nationaux. Par contre, ces métriques sont actuellement en perte de vitesse car elles ne sont pas adaptées à une évaluation individuelle. En effet, il a été montré qu'environ 15% des articles reçoivent 85% des citations au sein d'une même revue [9]. Pour plus d'objectivité au niveau individuel, il est plus pertinent de se baser sur des mesures d'impact au niveau des travaux des auteurs.

Les indicateurs bibliométriques directs les plus utilisés sont le h-index et le g-index.

En 2005, Jorge E. Hirsch publie un article fondateur à propos d'un nouvel indicateur, le **h-index**, qui a reçu un excellent accueil de la communauté scientifique : Il permet de prendre en compte l'impact précis des travaux d'un chercheur et d'évaluer sa carrière. Selon Hirsch, un scientifique possède un index h si h de ses publications ont reçu au moins h citations chacune et que ces autres publications ont reçu h citations ou moins [10].



H-index - Source : Wikipédia.

En 2006, Leo Egghe a proposé une variante du h-index, le **g-index** par laquelle il tente de corriger un défaut du h . En effet, ce dernier ne tient pas compte des citations reçues par les articles dont le rang dépasse la valeur h . La valeur g est définie comme suit: un scientifique possède un index g si ses g publications les plus citées ont reçu ensemble au moins g^2 citations [11]. Le g -index, qui est mathématiquement toujours plus grand ou égal au h -index, permet notamment de mieux différencier les chercheurs ayant peu de publications à leur actif mais dont les impacts en terme de nombre de citations reçues sont différents. Le ratio entre le g -index et le h -index est aussi intéressant car il donne une idée de la distribution des citations.

Afin d'évaluer au mieux l'ensemble du travail d'un chercheur, les métriques directes et indirectes sont en général **pondérées** en fonction du type de publication : par exemple on peut définir que les éditoriaux, lettres, revues de littérature et les études

[9] Adam D. Citation analysis: The counting house. Nature. [10.1038/415726a]. 2002;415(6873):726-9.

[10] Hirsch JE. An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(46):16569-72.

[11] Egghe L. Theory and practise of the g-index. Scientometrics. 2006;69(1):131-52.

de cas ont moins d'importance que les articles originaux. En général, le crédit que reçoit un auteur pour chacune de ses publications est aussi pondéré en fonction de la part de sa contribution. Cette dernière est souvent estimée par sa position dans la liste des auteurs. Souvent, le premier auteur est considéré comme le plus important. Cependant, selon les domaines, c'est le dernier auteur qui a la responsabilité de la publication. Mais encore, dans certaines branches comme la physique des particules, la tradition veut que les auteurs soient triés par ordre alphabétique. Au final, il faut savoir appliquer un modèle mathématique approprié à chaque discipline.

Pour une analyse valable, il est évidemment nécessaire de se baser sur des bibliographies exhaustives et contrôlées. De plus, pour chaque référence, il faut obtenir les données brutes sur chaque publication, notamment le compte des citations reçues. Or, ces informations sont répertoriées dans différentes bases de données, gratuites ou accessibles par abonnement:

Web of Science (WoS) [12]

Produite par Thomson Reuters, cette base de données enregistre les métadonnées et l'ensemble de citations figurant dans la bibliographie de chaque article publié dans la sélection des revues du JCR (environ 10'000 titres). C'est historiquement la première base de données contenant des données bibliométriques et elle demeure la plus vaste et la plus utilisée dans ce domaine. WoS contient à ce jour environ 46 millions de références primaires et plus de 750 millions de références citées, depuis 1900 à nos jours [13]. Cette base est constituée principalement de deux grands ensembles thématiques: Science Citation Index et Social Science Citation Index, complétés avec un jeu sur les articles de conférences.

SCOPUS [14]

En 2004, Elsevier lançait SCOPUS, une énorme base de données avec des fonctionnalités bibliométriques concurrençant Web of Science. SCOPUS indexe le contenu de 18'000 titres (y compris 1'200 journaux Open Access), 350 collections éditoriales de monographies et 3.6 millions d'actes de conférence, soit environ 40 million de références au total. Malheureusement, l'information bibliométrique concerne uniquement les articles publiés depuis 1996 (20 millions). L'autre moitié des références remonte à 1823, mais ne contient pas d'informations sur les citations [15].

PubMed Central (PMC) [16]

La National Library of Medicine (NLM) des Etats-Unis est pionnière dans le développement de bases de données A&I (Abstracting and Indexing) et des interfaces web de recherche associées. A côté du célèbre PubMed, la NLM a mis en place l'archive ouverte PubMed Central (PMC) en février 2000. Dans PMC, plus de

[12] <http://www.isiknowledge.com>

[13] Thomson Reuters. Web of Knowledge Quality and Quantity: Real facts, Real numbers, Real Knowledge. <http://wokinfo.com/realfacts/qualityandquantity/>

[14] <http://www.scopus.com>

[15] Elsevier. SCOPUS Content Coverage Guide. <http://info.scopus.com/scopus-in-detail/facts/>

[16] <http://www.pubmedcentral.nih.gov>

deux millions de notices sont enrichies avec des données de citations exploitables en bibliométrie.

Citebase [17]

Citebase est une base de données collectant des informations bibliométriques et des statistiques d'utilisation de la littérature scientifique publiée en Open Access. Cette base est alimentée par moissonnage automatisé des serveurs de pré- et post-prints ainsi que de dépôts institutionnels. Citebase extrait les références contenues dans la bibliographie des articles et indexe toutes ces métadonnées dans un moteur de recherche. Cet outil couvre surtout les domaines de la physique, des mathématiques, des sciences de l'information ainsi que des articles biomédicaux publiés en Open Access (BioMed Central ou PubMed Central). Citebase est exploitée par arXiv.org pour produire ses métriques de citations et d'usage.

CiteSeerx [18]

Cette bibliothèque numérique est spécialisée en informatique et en sciences de l'information. Il s'agit aussi d'une plateforme expérimentale et innovante, proposant de nouvelles méthodes d'analyse de citations ainsi que des algorithmes d'extraction de références bibliographiques à partir des documents PDF diffusés sur le web. CiteSeerX est capable de récupérer le nombre de citations, mais aussi les identifiants des documents citants.

Google Scholar [19]

Google Scholar pourrait représenter une alternative sérieuse aux outils classiques Web of Science et SCOPUS. Malheureusement, son fonctionnement n'est pas documenté et ses données, extraites automatiquement à partir de sources hétérogènes, sont peu fiables [20]. Malgré la forte demande, Scholar ne propose pas de service web et il faut donc extraire les informations bibliométriques directement depuis son interface web. Alternativement, il est possible d'utiliser des plugins du navigateur Firefox tel que Scholarometer [21] ou Publish or Perish (PoP) [22], permettant de comparer le h-index de Web of Science et SCOPUS avec celui de Google Scholar.

CrossRef [23]

L'organisation CrossRef vise à devenir la plaque tournante permettant d'accéder à l'ensemble de l'information académique en format électronique. Cette base de données de « Digital Object Identifiers » (DOI) forme l'une des plus grandes collections de métadonnées bibliographiques au monde (45 millions de DOIs enregistrés à ce jour). Parallèlement aux références bibliographiques, fournies directement par les éditeurs, CrossRef possède un système de connexion permettant

[17] <http://www.citebase.org>

[18] <http://citeseerx.ist.psu.edu>

[19] <http://scholar.google.com>

[20] Voir le cas de l'auteur fictif « Ike Antkare » : <http://www.enssib.fr/breves/tags/ike-antkare>

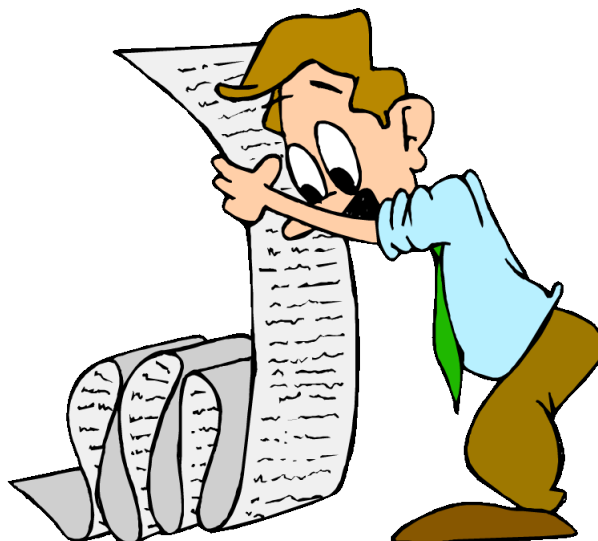
[21] <http://scholarometer.indiana.edu/>

[22] Harzing AW. Publish or Perish. <http://www.harzing.com/pop.htm>

[23] <http://www.crossref.org>

aux éditeurs de faire des liens hypertextuels entre les documents électroniques. Grâce à ces liens, les usagers peuvent aujourd'hui sauter directement d'une citation bibliographique vers le texte intégral du document associé. En 2004, CrossRef et l'entreprise Atypon ont lancé « CrossRef forward linking » [24], un service proposant aux éditeurs membres de CrossRef de déterminer le nombre de citations de leurs publications et de l'afficher directement sur leur plateforme de publication électronique. Ces informations de citation sont confinées aux sites web des éditeurs mais dans le futur CrossRef pourrait s'en servir pour concurrencer Web of Science et SCOPUS.

Comme vous pouvez le constater à travers cet échantillon de bases de données, le répertoire des sources prises en compte pour établir les métriques est très varié. La plupart des bases se limitent aux publications revues par le pairs (« peer reviewed ») et majoritairement en anglais ce qui exclut de la bibliométrie toute une série de publications et de types de documents qui se trouvent en dehors de ces frontières, pénalisant ainsi une partie de la production scientifique et de la vulgarisation. Pour palier quelque peu à cette limitation du périmètre de chaque base de données et à leurs lacunes respectives [25], il semble pertinent de combiner un maximum de sources adéquates, afin d'élargir le spectre des documents « citants » pris en compte [26]. Cependant, cette combinaison doit se faire avec prudence en se limitant uniquement aux sources dont le périmètre est connu et dont les données sont dignes de confiance. On obtiendra ainsi des résultats reproductibles et vérifiables.



Deux grands défis attendent la bibliométrie dans les années à venir. D'une part, la différenciation des citations selon une typologie précise et, d'autre part, l'amélioration des méthodes d'extraction automatiques des citations et la mise à disposition des informations bibliométriques basées sur les archives ouvertes.

En effet, si les auteurs et les éditeurs appliquaient dans les bibliographies une typologie unifiée permettant de différencier les citations, il serait possible de les distinguer et de les pondérer automatiquement d'une manière plus juste. Par exemple, en évitant d'attribuer un impact « positif » aux travaux cités dans un esprit

[24] http://www.crossref.org/02publishers/forward_linking_howto.html

[25] Jacsó P. Database source coverage: hypes, vital signs and reality checks. Online Inf Rev. 2009;33(5):997-1007; Adam D. Citation analysis: The counting house. Nature. [10.1038/415726a]. 2002;415(6873):726-9.

[26] Godel Spring S., De Kaenel I., Iriarte P. Web Services for Bibliometrics. EAHIL 2010 : 12th European Conference of Medical and Health Libraries : Discovering new seas of knowledge. - Lisboa, Portugal, 14th - 18th June 2010. http://archivesic.ccsd.cnrs.fr/sic_00540289/

critique. L'initiative CiTO (Citation Typing Ontology) [27] propose un système permettant de faire cette catégorisation ontologique et de coder cette information dans les métadonnées du document afin qu'elle puisse être exploitée par des techniques sémantiques et bibliométriques.

L'autre défi majeur consiste à améliorer les méthodes d'extraction automatisée des citations à partir du texte intégral des documents déposés dans les archives ouvertes. En effet, ces techniques sont déjà utilisées et permettent par exemple à PubMed Central d'extraire et d'exploiter ces informations selon deux axes complémentaires : la bibliométrie avec des indicateurs calculés au sein du corpus de l'archive, et des données sémantiques décrivant les liens entre les documents. Ces deux types d'informations pourraient devenir plus fréquemment exploitables si de nouveaux dépôts thématiques et institutionnels se mettaient à les calculer et à les rendre disponibles pour la communauté comme le fait PubMed Central. Cela permettrait d'élargir le corpus des documents sources utilisé pour le calcul de l'impact d'une publication.

En conclusion, nous sommes convaincus que les bibliothécaires, avec leur expérience dans le contrôle et la validation des informations biblio-graphiques, peuvent apporter un grand plus à la bibliométrie. Et c'est une opportunité à saisir! Nous nous sommes donc lancés dans le développement d'un logiciel libre généraliste d'analyse bibliométrique suffisamment flexible pour être capable de répondre aux besoins des institutions académiques, et en particulier de ceux des Universités de Lausanne (UNIL) et de Genève (UNIGE).



[27] Shotton, D. CiTO, the Citation Typing Ontology. Journal of Biomedical Semantics 2010, 1(Suppl 1):S6. doi:10.1186/2041-1480-1-S1-S6. <http://purl.org/net/cito/>