Les citations ouvertes

Analyse I/IST-n°28-Septembre 2018

Les citations sont un instrument privilégié de la recherche bibliographique : le repérage des citations ultérieures d'une publication permet de reconstituer rapidement l'état d'un champ de recherche ou la réception d'une découverte ou d'une théorie scientifique.

Jusqu'en 2017, ces informations étaient concentrées dans de grandes bases fermées telles que le *Web of Science*, *Scopus* (d'Elsevier) ou plus récemment *Google Scholar* (de Google)[1] et collectées selon des critères opaques. En un an tout a changé : suite au lancement d'une Initiative pour les citations ouvertes (I4OC) la part des données de citation mises à disposition sous une licence libre est passée de 1% à 51%[2] des références scientifiques disponibles sous Crossref. Ces données sont mises à disposition en totalité sur la plateforme OpenCitations et peuvent être massivement réutilisées sur d'autres projets comme Wikidata.

L'ouverture des citations s'inscrit dans une nouvelle dynamique du libre accès qui s'étend désormais bien au-delà de la simple mise à disposition des écrits scientifiques pour intégrer de nouvelles formes : données, code, évaluations, métadonnées… L'enjeu n'est plus seulement de faciliter l'accès à la connaissance mais aussi de transformer radicalement les structures de publication et de dissémination de la recherche.

Un premier essai avorté : l'index de citation ouvert d'Eugene Garfield.

L'ouverture des données de citation a bien failli avoir lieu 50 ans plus tôt.

À la fin des années 1950, la revue scientifique semble condamnée à disparaître. Face à un essor sans précédent du nombre de publications, le gouvernement américain envisage de développer une infrastructure centralisée reposant sur des « dépôts » d'articles : il était notamment déjà question de numériser la totalité des productions scientifiques pour un coût annuel total de 10 000 000 $[3].

La mise en relation entre tous ces documents désormais dissociés du format périodique serait assurée par plusieurs instruments de recherche dont un index de citation alors confié à Eugene Garfield[4], mais aussi des bases de données électroniques permettant des recherches par mots et des « méta-périodiques » diffusant quotidiennement des résumés et des comptes rendus des publications les plus intéressantes.

Faute de financement, seuls quelques éléments de ce vaste programme verront le jour. La base de données *MEDLINE* est sans doute l'héritier le plus direct : constituée en 1964 et diffusée en ligne dès 1971, elle permet d'interroger les métadonnées des articles de 2500 revues de médecine. De son côté, Garfield parvient à réaliser l'index de citation mais sous la forme d'une entreprise commerciale privée et non d'une organisation « non-for profit » comme initialement envisagé. En 1963, son *Institut de l'information scientifique* fait paraître le *Science Citation Index* combinant les références utilisées dans quelques milliers de revues de référence. Désormais coupée du « dépôt institutionnel », la base de données bibliométrique va fortement encourager le développement d'un autre modèle : la grande revue internationale contrôlée par de grands éditeurs. Le Science

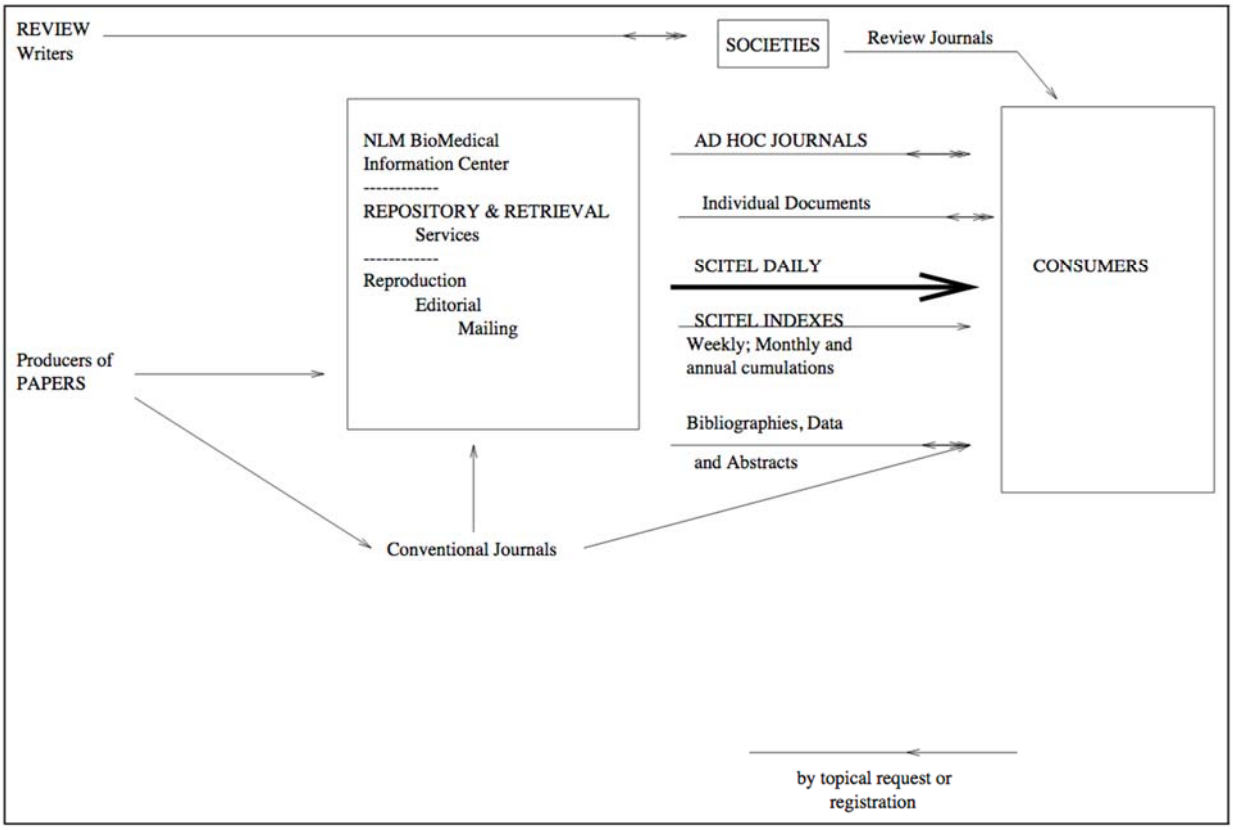
[1] Si Google Scholar est accessible en libre accès, les données de citation sont quant à elles protégées et le téléchargement automatique est fortement découragé (https://scoms.hypotheses.org/216)

[2] https://i4oc.org/news.html#April2018

[3] Sur ce contexte historique, voir la thèse de Paul Wouters, *The citation culture*, 1999, https://dare.uva.nl/search?identifier=b101b769-100f-43e5-b8d2-cac6c11e5bbf.

[4] Garfield avait émis une première proposition théorique d'un index de citation dès 1955 : Eugene Garfield, « Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas ». In : *Science*. 15 juillet 1955. Vol. 122, n° 3159, p. 108-111.

Citation Index devenu aujourd'hui Web of Science se focalise progressivement sur le facteur d'impact et devient un instrument d'une publication ou d'une carrière. Après avoir longtemps envisagé d'acheter l'index de Garfield, le premier éditeur mondial Elsevier lance en 1997 sa propre plateforme collectant les citations, Scopus.



L'infrastructure de dépôt centralisé projetée par Joshua Lederberg en 1962[5].

L'intégration de l'index de Garfield dans un programme fédéral aurait sans doute eu pour effet de le transformer en base ouverte. Dans la loi américaine, les productions de l'État fédéral relèvent en effet du domaine public : en tant que projet mis à disposition par la National Library of Medicine, MEDLINE constituait ainsi un recueil « d'informations dans le domaine public qui peut être librement distribué et copié[6] ».

L'impulsion originelle : des sources pour Wikipédia!

L'index de citation ouvert mis aujourd'hui à disposition par *OpenCitations* est né dans un environnement radicalement différent : non pas dans la continuité d'un grand projet d'infrastructure publique mais suite aux besoins spécifiques d'un projet communautaire autogéré, né hors du monde académique, Wikipédia.

Vers 2004-2005, l'encyclopédie collaborative prend de l'ampleur et afin de gagner en légitimité entreprend de généraliser l'emploi des références. La « vérifiabilité » devient le critère fondamental de la fiabilité d'une information : dans la mesure où Wikipédia ne produit aucun savoir mais représente le savoir existant, les articles doivent être nécessairement élaborés à partir de sources fiables et dûment citées[7]. L'infrastructure évolue pour intégrer les notes de bas de page ou des bibliographies sous forme de données structurées. Plusieurs contributeurs constatent alors qu'une grande base de données bibliographiques simplifierait significativement ce travail de documentation. Au lieu de devoir spécifier à chaque fois la source manuellement,

[5] Schéma reproduit par Paul Wouters dans sa thèse précitée, p. 68.

[6] https://www.nlm.nih.gov/copyright.html

[7] Sur cette évolution, voir mon étude : « {{Référence nécessaire}} : l'émergence d'une norme wikipédienne (2003-2009) », dans Wikipédia et la Science (sous la direction de Valérie Schaeffer, Lionel Barbe et Louise Merzeau) : Presses universitaires de Paris-Ouest, 2015, p. 77-90. http://books.openedition.org/pupo/4106

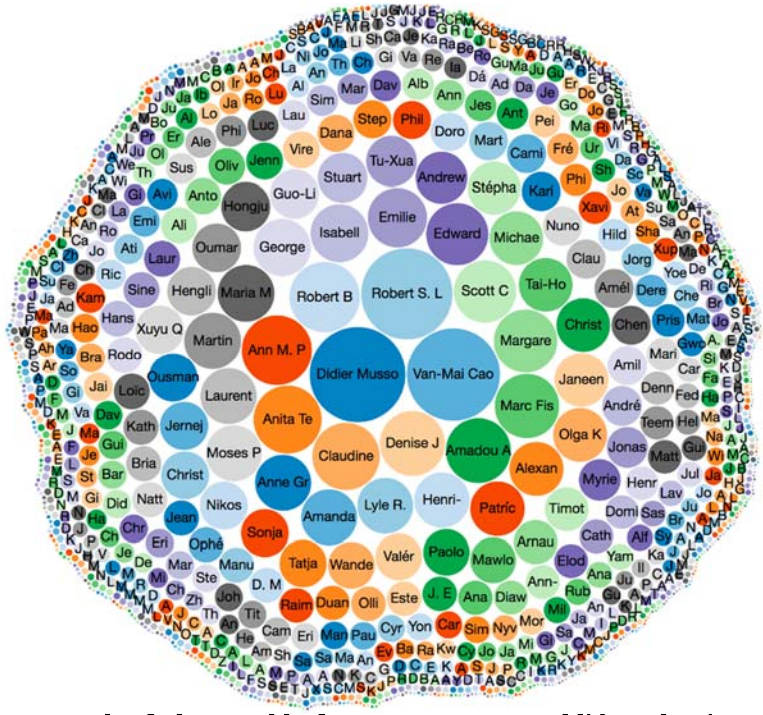
il serait possible d'importer rapidement une citation à partir d'un titre ou d'un identifiant pérenne (ISBN, ISSN, doi…). C'est ainsi que naît en 2006 le projet Wikicite qui, dans sa première version ambitionne de développer « une base de données séparée contenant les données de citation que l'on pourrait extraire pour générer une référence complète et standardisée à la fin de l'article[8]. »

Bien que périodiquement évoqué dans les années suivantes, cet idéal n'est jamais concrétisé. L'émergence de Wikidata en 2013 change la donne. Ce « Wikipédia des données » permet de créer manuellement ou automatiquement des millions de fiches sur tous les champs de la connaissance. Wikidata est naturellement conçu pour héberger des données bibliographiques. C'est même un objectif prioritaire : tout comme Wikipédia, Wikidata repose sur un principe de vérifiabilité qui veut que chaque donnée soit référencée ; tout comme sur Wikipédia, la mise en œuvre de ce programme est compliquée par la nécessité de remplir la fiche de chaque nouvelle référence.

Le projet *Source Metadata* est initialement créé en 2014 pour établir des standards bibliographiques. Il réalise finalement l'objectif originel de Wikicite en développant une « large base de données bibliographiques sur Wikidata ». Les contributeurs procèdent à l'importation préventive massive de références grâce à l'emploi de robots. Actuellement plus de 20 millions de sources sont déjà présentes sur le projet.

Cette intégration systématique constitue actuellement un défi pour l'ensemble de Wikidata : 40% des items (les « fiches » correspondant aux articles de Wikipédia) sont des références bibliographiques et cette part va sans doute continuer de croître. En août 2018, une discussion a été ouverte pour statuer sur l'avenir de cette base de données bibliographiques ouverte avec notamment la possibilité de créer un statut à part pour ces objets, voire de créer un nouveau projet[9].

Entretemps le projet *Wikicite* s'est structuré au-delà de Wikidata, notamment suite à l'organisation d'une première conférence à Berlin en 2016, depuis transformée en rendez-vous annuel. C'est lors de la conférence qu'est créée la propriété « cite », qui permet d'enregistrer les citations de chaque référence sur Wikidata. Initialement contestée par une partie de la communauté lors de sa création[10], elle est aujourd'hui la 16e propriété[11] la plus utilisée sur Wikidata avec plus de 4 millions d'usages. *Wikicite* contribue également au lancement de projets « pilote » tel que le recensement de l'ensemble des publications traitant du virus Zika, alors à l'origine d'une épidémie virulente au Brésil.



Graphe de l'ensemble des auteurs ayant publié sur le Virus Zika, utilisé comme exemple des avancées des travaux de Wikicite[1].

L'émergence rapide d'un index ouvert

Quelques mois plus tard, Dario Tarobelli présente les acquis du projet *Wikicite* lors de la conférence annuelle

---

[8] https://meta.wikimedia.org/wiki/Wikicite_(metadata_proposal)

[9] https://www.wikidata.org/wiki/Wikidata:WikiCite/Roadmap

[10] Des inquiétudes se faisaient déjà jour sur le risque « d'éclater Wikidata avec des millions de documents citant chacun des dizaines voire des centaines de documents ». https://www.wikidata.org/wiki/Wikidata:Property_proposal/cites

[11] https://www.wikidata.org/wiki/Wikidata:Database_reports/List_of_properties/all

des éditeurs de libre accès (COASP)[12]. À la fin de son intervention, il présente deux requêtes de la communauté : ouvrir les données de citation et autoriser l'extraction automatisée de contenu. L'impact est immédiat. Un petit groupe informel de réflexion se constitue dans la foulée et formera l'embryon de l'Initiative pour les citations ouvertes.

L'écosystème de la citation ouverte se structure alors autour de quatre acteurs :

- L'**Initiative pour les citations ouvertes** (I4OC)[13] milite pour l'ouverture des données.
- L'agence **Crossref**[14], qui gère notamment l'attribution des identifiants d'articles, les DOI, réceptionne les citations mises à disposition pour les éditeurs.
- Le projet **OpenCitations**[15] diffuse un corpus sous licence CC0 de 316 millions de citations de 24 millions de publications[16]. Originellement, il s'agit d'un projet scientifique beaucoup plus modeste, financé par le JISC dès 2010 : la première version du corpus publiée en 2016 comprenait quelques centaines de milliers de citations, soit mille fois moins qu'aujourd'hui[17]. Le corpus d'OpenCitations est actualisé tous les six mois, notamment à partir des informations disponibles dans Crossref. La dernière version date de juillet 2018.
- **Wikidata** réutilise en partie ces données pour enrichir les fiches des références bibliographiques et les met donc en relation avec une vaste base de connaissance. Actuellement, le projet dispose d'un corpus de 100 millions de citations réparties sur 20 millions d'article[18].

Cette gouvernance polycentrique entre étroitement en résonance avec le processus de diffusion et d'entretien d'une ressource en libre accès et contraste avec les structures hiérarchiques des index fermés.

Dès le lancement officiel de l'Initiative pour les citations ouvertes, une vingtaine d'éditeurs sont déjà associés et se sont engagés à mettre à disposition les citations de 40% des références disponibles sur Crossref. Trois mois plus tard, en juillet, cette part s'élève à 45%[19]. Lors des derniers calculs effectués par l'Initiative en avril 2018, plus de la moitié des références de Crossref étaient concernées (soit 51%)[20].

Le ralentissement de la croissance après l'accélération initiale est surtout imputable à l'absence d'Elsevier, qui concentre 25-30% des citations. De toute évidence, le premier éditeur mondial ne tient pas à libérer l'accès des données qu'il valorise directement sur Scopus. Une timide dynamique d'ouverture s'esquisse tout de même : dans la dernière version du corpus d'OpenCitations, près de 2 millions de citations proviennent d'Elsevier, même si ça ne représente encore qu'une part très minoritaire des données qui restent à ouvrir (probablement de l'ordre de 100 millions).

L'essor rapide des citations ouvertes hors Elsevier s'explique en grande partie par la conversion déjà réalisée d'une partie de l'édition scientifique au libre accès : l'appropriation des données de citation ne présente plus d'intérêt commercial réel dans ce contexte, voire contribue à pérenniser la concentration oligopolistique du milieu. Le Web of Science et Scopus reposent sur une conception restrictive de la publication scientifique qui discrimine des formats très utilisés dans certaines disciplines (chapitre, ouvrage, compte rendu…) et favorise les grandes revues historiques.

---

12 https://figshare.com/articles/Citations_needed_for_the_sum_of_all_human_knowledge_Wikidata_as_the_missing_link_between_scholarly_publishing_and_linked_open_data/3956238

13 https://i4oc.org/

14 https://www.crossref.org/

15 http://opencitations.net/

16 http://opencitations.net/download La licence CC0 est globalement équivalente au domaine public et retire toutes les restrictions existantes dans la mesure où la loi le permet (ce qui pour les données revient notamment à ne pas rendre l'attribution obligatoire).

17 Silvio Peroni, David M. Shotton & Fabio Vitali, *Freedom for bibliographic references: OpenCitations arise*, 2016, https://www.semanticscholar.org/paper/Freedom-for-bibliographic-references%3A-OpenCitations-Peroni-Shotton/3ba6238f9cf460e6e7d93d240f7662c9368b9cf0

18 http://wikicite.org/statistics.html

19 https://elifesciences.org/for-the-press/9eb16f1c/availability-of-open-reference-data-nears-50-as-major-societies-and-influential-publishers-endorse-the-initiative-for-open-citations

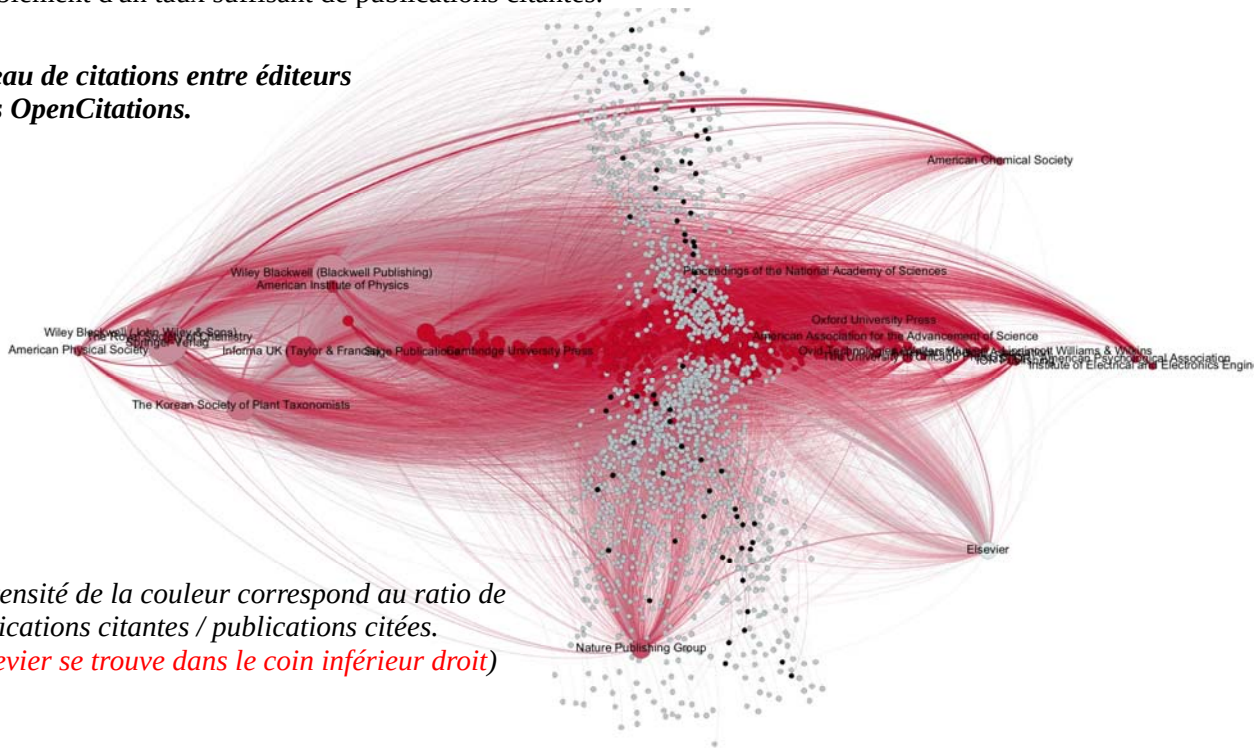
20 https://github.com/elifesciences/datacapsule-crossref/blob/analysis/notebooks/citation-stats.ipynb

Les éditeurs français dans l'index ouvert des citations.

Le corpus complet actuellement mis à disposition par OpenCitations a été constitué en juillet 2018, à partir d'une extraction de Crossref. Le principal jeu de donnée, "Citation data" enregistre 316 millions de couples de publications citantes et de publications citées, ainsi que quelques métadonnées contextuelles (date de publication, délai après la parution de la publication citée…). Nous avons récupéré l'ensemble de ces données et avons procédé à l'extraction et la réidentification du préfixe du DOI qui intègre le nom de l'éditeur (par exemple Open Edition est 10.4000). Sur R cette opération prend environ une heure.

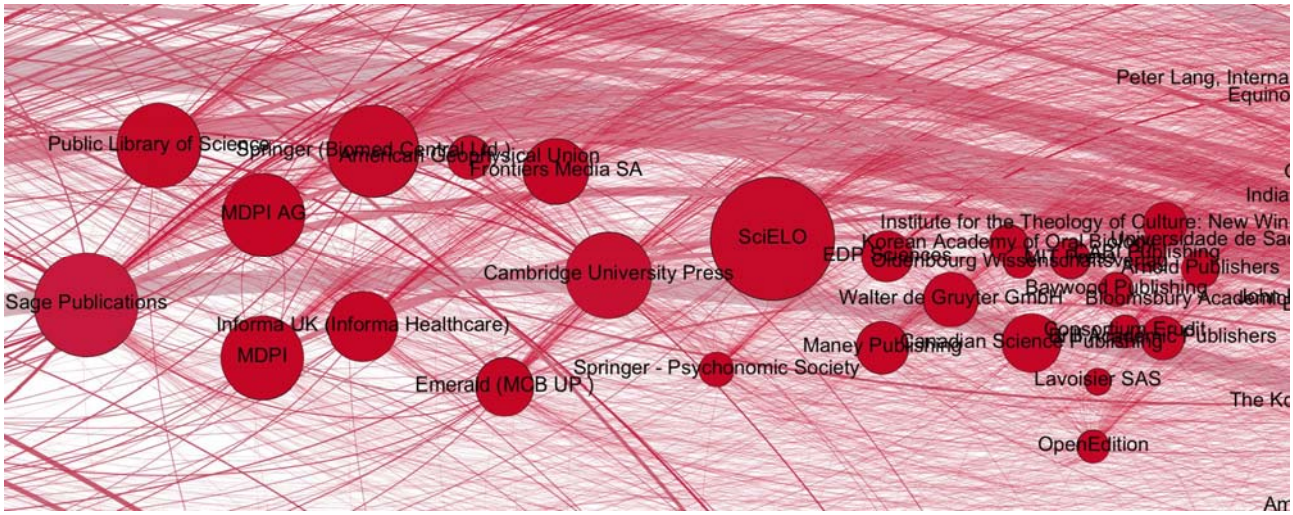
En raison de la faible participation de plusieurs éditeurs de premier plan, l'index renverse la cartographie usuelle de l'édition scientifique. Tout en restant quantitativement l'éditeur le plus cité (avec 76 millions de citations), Elsevier se retrouve en position périphérique dans notre projection en réseau des éditeurs, faute probablement d'un taux suffisant de publications citantes.

Réseau de citations entre éditeurs dans OpenCitations.



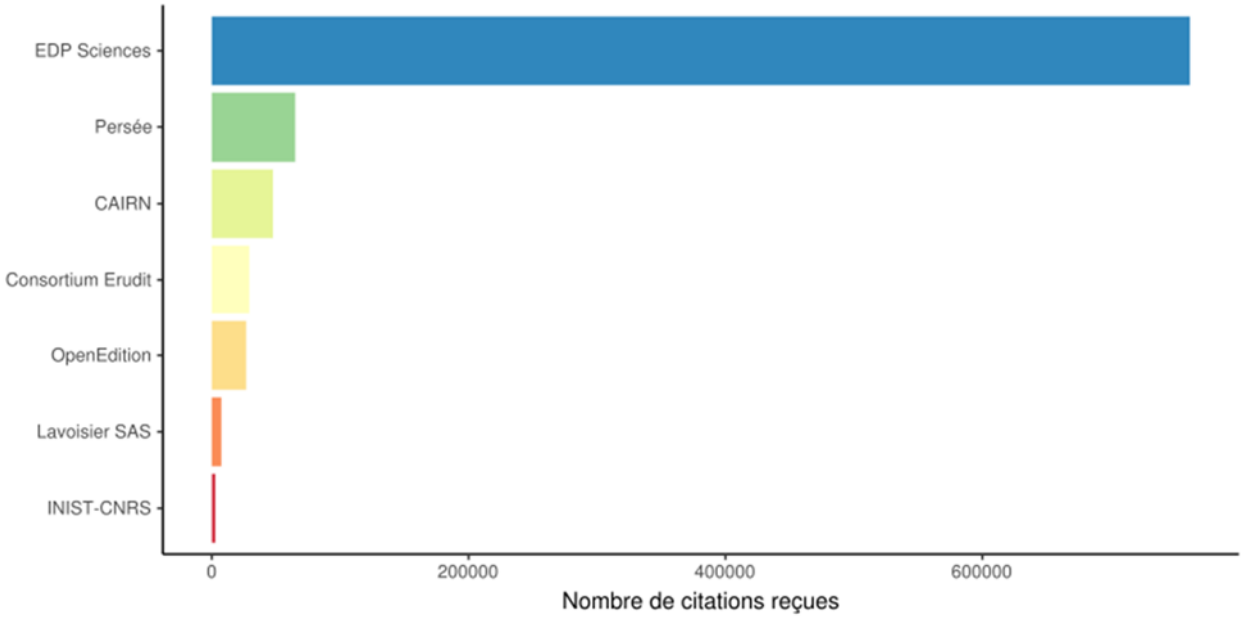
L'intensité de la couleur correspond au ratio de publications citantes / publications citées.
(Elsevier se trouve dans le coin inférieur droit)

Par compensation, le cœur du réseau est constitué d'organisations de taille moyenne souvent impliquées de longue date dans le mouvement du libre accès : PLOS, Scielo, EDP Sciences, Open Edition. Cette place centrale illustre une transformation majeure de l'édition scientifique, l'oligopole des grands éditeurs déclinant au profit de nouveaux entrants.



Zoom sur le cœur du réseau, structuré autour d'acteurs historiques du libre accès.
Open Edition se trouve tout en bas à droite.

L'index de citation propose une couverture bien supérieure de disciplines jusqu'ici négligées par le Web of Science ou par Scopus. Des éditeurs francophones majeurs de sciences humaines et sociales comme Open Edition (168 000 citations) ou le consortium Erudit (128 000 citations) participent à l'initiative, ce qui contribue à représenter d'autres acteurs du champ (CAIRN, Persée, les éditions CNRS…) parmi les publications citées.



Les principaux éditeurs français par nombre de citations reçues dans OpenCitations

L'index prend aussi en compte des livres et recueils de conférences : dans sa dernière version Citation Data comprend 2741 citations en provenance d'Open Edition Books. Pour l'instant, cet élargissement reste encore pénalisé par une faible indexation de ces productions sur Crossref. Le taux de citation par chapitre d'ouvrage d'Open Edition Books est anormalement bas, avec 1,07 citation en moyenne pour les chapitres ayant déclaré au moins une citation, contre 21 citations en moyenne pour toutes les publications référencées dans Citation Data. Les auteurs citent vraisemblablement des ressources (livres ou chapitres publiés chez d'autres éditeurs) qui n'ont pas d'identifiant enregistré dans Crossref.

Cette première analyse inspire plusieurs observations. En raison d'une émergence très rapide, l'écosystème de la citation ouverte demeure incomplet. Plusieurs données bibliographiques essentielles font encore défaut pour procéder à certaines analyses. Il n'existe apparemment pas de base à jour des "préfixes" de DOI, qui contiennent l'identifiant de l'éditeur[21]. Crossref ne collecte pas d'informations sur l'ancrage disciplinaire des revues (ce qui permettrait par exemple d'établir un « taux de transdisciplinarité » à partir des citations croisées). Par ailleurs, sous sa forme actuelle, le jeu de données est assez volumineux et va sans doute le devenir davantage. L'objectif actuel de couverture de 100% des références de Crossref représenterait plus de 600 millions de citations et ce chiffre reste bien en deçà de la totalité des citations documentées dans les publications scientifiques. Pour rester maniable, le corpus devra probablement se décliner par date ou par éditeur.

Un écosystème à construire

L'ouverture des citations contribue partiellement à combler une « brique » manquante de l'édition scientifique en libre accès : l'indexation des publications. Jusqu'à présent, les grands acteurs de l'édition scientifique n'avaient été concurrencés sur ce plan que par des entreprises et start-up guère plus ouvertes, telles que les réseaux sociaux académiques. Par contraste, le mouvement initié par l'Initiative pour les citations ouvertes s'appuie sur une convergence inédite entre communautés scientifiques, institutions publiques et « commun » autogéré.

Pour l'heure, une part significative de données de citation sont ouvertes ; elles ne sont pas encore accessibles. Tout un écosystème reste à construire en interaction étroite avec bibliothèques numériques, plateformes de

[21] Ici, nous avons récupéré les données mises à disposition sur Wikidata par d'autres projets plus anciens. Nous en avons aussi manuellement complété certaines.

publications en libre accès et moteur de recherche spécialisé (comme Isidore en France). Une application comme Scholia donne déjà un aperçu de ce qui peut être réalisé : elle permet de naviguer entre les nombreuses statistiques disponibles pour un auteur, un article ou un thème sur Wikidata et de les positionner ainsi par rapport à des champs de recherche ou des communautés scientifiques.

Exemples de statistiques générées par Scholia pour Peter Murray-Rust[22] : graphe des co-auteurs,

Citing authors

Authors that cite the author (excluding self citations).

Show 10 entries

Count	Citing author	Orcid
53	Egon Willighagen	0000-0001-7542-0286
31	Christoph Steinbeck	0000-0001-6966-0814
28	Ola Spjuth	0000-0002-8083-2864
25	Jeremy G. Frey	0000-0003-0842-4302
18	Rajarshi Guha	0000-0001-7403-8819
15	Marcus D. Hanwell	0000-0002-5851-5272
13	Jonathan Alvarsson	0000-0002-8682-7206
13	Andreas Bender	0000-0002-6683-7546
12	Michel Dumontier	0000-0003-4727-9435
11	Chris T. Evelo	0000-0002-5301-3142

thématiques traitées dans les publications, liste des principaux auteurs citants.

Au-delà de l'accès, de nombreuses données complémentaires restent encore à acquérir pour que les citations ouvertes deviennent pleinement exploitables. Les données bibliographiques de plus haut niveau restent encore insuffisantes et ne permettent pas encore d'analyser certains phénomènes significatifs comme les interactions entre disciplines. Dans certains domaines comme les sciences humaines et sociales, l'enjeu n'est pas tant d'obtenir des éditeurs l'ouverture des données que de les « créer » en premier lieu : livres, chapitres et comptes rendus devraient faire l'objet d'une extraction systématique[23]. Enfin, la mise à disposition du texte complet permet d'envisager de reconstituer le « contexte » de la citation avec des techniques de *text & data mining*, en identifiant par exemple les entités concernées par la citation (gène, composé chimique, personnalité, institution…) ou en repérant des *modalités citationnelles* (notamment les allusions critiques, voire polémiques, à des travaux antérieurs).

Sur tous ces aspects, la mise en œuvre d'une politique active de soutien aux infrastructures de recherche apparaît comme une nécessité. Plusieurs institutions majeures de la recherche européenne viennent de dévoiler, avec le soutien de la Commission, un plan audacieux de transition vers l'édition au libre accès (dit « plan S ») décliné en 10 principes. L'objectif n'est plus seulement de faire basculer les dépenses déjà engagées pour l'acquisition de publications fermées (« journal flipping ») mais bien de restructurer profondément l'édition scientifique en favorisant le développement de nouvelles plateformes, dès lors que l'offre actuelle fait défaut (3e principe)[24]. C'est exactement le type d'impulsion dont le nouvel écosystème des citations ouvertes aurait aujourd'hui besoin pour se développer pleinement.

[22] https://tools.wmflabs.org/scholia/author/Q908710

[23] Typiquement, la fiche Scholia de Michel Foucault ne contient actuellement aucune citation : https://tools.wmflabs.org/scholia/author/Q44272

[24] https://www.scienceeurope.org/coalition-s/