HAL
archives-ouvertes.fr

Le millefeuille des affiliations françaises dans les publications scientifiques

Michèle Dassa, Alina Deniau, Manuel Durand-Barthez, Françoise Girard, Nathalie Pothier, Angèle Séné

To cite this version:

Michèle Dassa, Alina Deniau, Manuel Durand-Barthez, Françoise Girard, Nathalie Pothier, et al.. Le millefeuille des affiliations françaises dans les publications scientifiques. Documentaliste - Sciences de l'Information, ADBS, 2014, Le Droit sans complexe : décryptage et repères, 51 (4), pp.12-16. http://www.adbs.fr/b-methodes-techniques-et-outils-b-br-le-millefeuille-des-affiliations-francaises-br-dans-les-publications-scientifiques-145188.htm?RH=REVUE . 10.3917/docsi.514.0012 . sic_01097580

HAL Id: sic_01097580
https://archivesic.ccsd.cnrs.fr/sic_01097580

Submitted on 19 Dec 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Article « **Le millefeuille des affiliations françaises dans les publications scientifiques** », par le groupe de travail « Affiliations » Renatis-Urfist : cette version auteur a été acceptée pour publication dans la revue *Documentaliste – Sciences de l'information*, dans son n° 4/2014.

[**recherche scientifique**] À l'heure où l'évaluation et le financement sous-jacent de la recherche sont en jeu, l'affiliation des acteurs de la recherche publique française est une question essentielle. Le groupe Renatis/Urfist présente les conclusions de ses travaux.

Le millefeuille des affiliations françaises dans les publications scientifiques

Le paysage scientifique français, en constant changement, se complexifie avec la multiplication des structures de recherche. Avec l'impact croissant des grands classements internationaux et des nouveaux modes d'évaluation et de financement, renforcer la visibilité scientifique française au niveau international est au cœur de tous les débats.

Contexte

Calqués sur le modèle anglo-saxon de la structuration de la recherche, les classements internationaux (Shanghai, Leiden, U-Multirank, QS-Ranking, etc.) tendent à privilégier les universités. La France, où la recherche est aussi conduite dans les organismes de recherche proprement dits, mal pris en compte dans ces classements, est de ce fait pénalisée.

Ce sont des bases de données bibliométriques, comme le Web of Science (WoS)[1] ou Scopus, qui mesurent les impacts des entités comparées, soit par le biais d'indicateurs dédiés et / ou par les classements internationaux. Cette visibilité passe par la pierre angulaire que constitue l'affiliation dans les publications, jalon incontournable de la production

[1] WoS agréé notamment de façon exclusive par la LOLF (Loi organique relative aux lois de finances).

scientifique. La notoriété des organismes, établissements, laboratoires, etc., dépend donc du fonctionnement et de la structuration des bases de données commerciales de réputation internationale interrogées pour élaborer ces classements ou indicateurs.

Associé à l'URFIST[2] de Paris dans le cadre de son atelier de Bibliométrie, le Réseau Renatis[3], réunissant des professionnels de l'information scientifique et technique, a entrepris en 2013 une réflexion nationale partant d'une étude relative aux pratiques de signature des publications scientifiques des laboratoires affiliés au CNRS en Midi-Pyrénées et Languedoc-Roussillon[4].

Problématique

Elle concerne la caractérisation, la valorisation, l'évaluation de la recherche.

Caractérisation et valorisation

Les professionnels de l'IST sont confrontés concrètement aux opérations de recensement de la production scientifique de leur institution notamment à des fins d'évaluation. Ces opérations sont délicates et rendues difficiles du fait de l'immense variété des intitulés d'affiliation mais aussi des limites et contraintes des bases de données actuelles. Abréviations, sigles, nombres, entités à géométrie variable, alternatives orthographiques, complexifient le travail de ces agents. Un calcul équitable ne pouvant se fonder que sur une base «propre», ils sont incités à élaborer des stratégies chronophages pour simuler et restituer toutes les modalités sémantiques imaginables. L'usage d'outils sophistiqués de retraitement de l'information, parfois coûteux et d'une ergonomie peu intuitive, est généralement indispensable. Or, temps, moyens financiers, matériels et humains font souvent défaut, *a fortiori* lorsque les délais impartis sont réduits.

Évaluation

La faible visibilité des universités et de la recherche française dans les classements internationaux a questionné la communauté scientifique française des chercheurs et des décideurs. Il est clair que, dans une logique de classement, le recensement des publications, à savoir des preuves patentes du travail effectué par les chercheurs, est capital. Si l'inventaire est effectué dans des conditions d'incomplétude notoire, trop souvent floues, l'évolution de carrière des chercheurs sera directement impactée. Le problème est donc loin d'être exclusivement technique et documentaire. Il est déontologiquement important, les conséquences d'un travail de collecte, d'inventaire et de dénombrement inaboutis se répercutant immanquablement sur l'appréciation qu'exerce une autorité administrative décisionnaire du travail fourni par les unités de recherche et leurs entités de tutelle.

---

[2] Urfist : Unité Régionale de Formation à l'Information Scientifique et Technique http://urfist.enc.sorbonne.fr/

[3] Renatis, réseau national de l'information scientifique et technique (IST) du CNRS http://renatis.cnrs.fr/

[4] Ce rapport (2012) de N. Roquefere, M. Villeneuve, et V. Vincent « *Affiliations des producteurs de la recherche publique française* », commandité par le Réseau Doccitanist, un des réseaux régionaux de Renatis, collationne de nombreuses circulaires sur la normalisation des graphies d'affiliations et propose une analyse critique des méthodes utilisées, comparant la théorie à la pratique http://doccitanist.lirmm.fr/spip.php?article223

Constat

À l'origine, les adresses des auteurs des articles scientifiques permettaient les échanges entre lecteurs et auteurs. Il ne s'agissait donc pas forcément d'appartenances institutionnelles. Avec l'accroissement mondial du volume des publications, les méthodes et outils d'évaluation de cette production ont pris des proportions importantes et les adresses, devenues affiliations professionnelles, sont les clés de base des études bibliométriques.

En principe une affiliation comprend au moins deux éléments : le laboratoire de l'auteur (nom et lieu) et sa tutelle (ou plus), c'est-à-dire l'établissement ou l'organisme auquel le laboratoire est rattaché. Le repérage des publications d'une entité de recherche repose surtout sur l'interrogation des affiliations correspondant à cette entité, ce qui implique la normalisation de ces adresses.

Trois faits nuisent à la visibilité de la France dans les classements au niveau international : l'hétérogénéité des adresses, l'aspect protéiforme des libellés des affiliations des publications scientifiques françaises et l'utilisation sans précaution des algorithmes commerciaux.

Hétérogénéité de l'écriture des affiliations

Connu de longue date, le manque de normalisation des affiliations provoque une déperdition importante de publications. Une étude a démontré que 44 % des publications de l'université Claude Bernard de Lyon référencées dans différentes bases de données ne mentionnent pas cette université[5]. De même, au CNRS, toutes disciplines confondues, cette absence de mention a été évaluée à hauteur d'environ 30 % dans le WoS.

En 2007, le groupe «Normadresses» piloté par l'OST[6] a proposé une harmonisation au niveau national des règles de signature[7]. Deux modèles, le mono-ligne et le multi-ligne, ont été comparés, avec leurs avantages et inconvénients. De fait, il n'a pas formulé de propositions communes et claires pour l'ensemble des acteurs de la recherche française mais uniquement des recommandations techniques de bon sens[8]. Par ailleurs, des chartes institutionnelles, construites en mode *top-down*, sont régulièrement promulguées indépendamment les unes des autres. Ces préconisations, parfois contradictoires entre les tutelles d'une même Unité mixte de recherche (UMR), inapplicables au niveau du chercheur, apportent encore plus de confusion et compliquent la construction d'indicateurs bibliométriques fiables.

Des adresses protéiformes

Le regroupement politique et stratégique d'entités de recherche, autre tentative pour remédier au problème de visibilité, se solde par un rallongement démesuré des lignes d'adresses dans les publications.

A priori simple, le schéma d'écriture des affiliations se complique au fur et à mesure de la multiplication des structures administratives. L'unité de base de l'organisation universitaire

[5] Pascal BADOR, Thierry LAFOUGE, La difficulté d'accéder aux adresses des chercheurs français dans les bases de données bibliographiques L'exemple d'une université lyonnaise. *Documentaliste - Sciences de l'Information*, 2006, 43 (1), p.28-35.

[6] Observatoire des Sciences et Techniques
http://urfist.enc.sorbonne.fr/sites/default/files/mdb/Normadresses_OST2007.pdf

[7] http://www.obs-ost.fr/sites/default/files/ReleveDecision_Normadresses_OST2007_01.pdf

[8] Cf. p. 22 du rapport *Normadresses*, URL ci-dessus note 6

française tend progressivement vers l'UMR, un laboratoire de recherche ayant au moins 2 tutelles, le plus souvent un établissement d'enseignement supérieur et un organisme de recherche. À ceci s'ajoute l'apparition de nouvelles structures géographiques et/ou politiques, qui peuvent de surcroît s'avérer éphémères, comme les PRES qui ont cédé leur place aux COMUE, ou tout simplement les structures de financement liées aux investissements d'avenir et d'excellence comme les LABEX, etc. De ce fait, le nombre d'informations à indiquer dans l'affiliation d'un auteur français pourrait augmenter progressivement. L'imbrication de structures en couches successives, tel un millefeuille, incluant pour une même entité des UMR, UPR[9], Pôles de compétitivité et/ou d'excellence etc., en constante recomposition, rend la notion d'appartenance institutionnelle de plus en plus confuse et illisible.

Impact des algorithmes

Retravaillées lors des opérations de la chaîne éditoriale, les retranscriptions de l'affiliation ne sont pas toujours fidèles au manuscrit du chercheur, ce qui a un impact sur les données finales exploitables à partir des bases commerciales. Par ailleurs, les bases de données ont leurs propres standards. Dans le WoS, par exemple, l'affiliation des auteurs d'une publication est systématiquement retraitée par un algorithme spécifique pour la normaliser et la rapprocher du modèle universitaire anglo-américain. Même si la publication est signée par plusieurs organismes, un seul est enregistré en tant que tutelle principale, le WoS se limitant à une unique «*Organization*». Cette mention est en priorité attribuée à l'université, quelle que soit sa place au sein de l'adresse de publication ou, par défaut, attribuée à l'organisme occupant la première place dans la ligne d'affiliation, ce qui a un effet direct sur l'élaboration d'indicateurs. Les organismes de recherche français, de type EPST[10] ou Epic[11], n'ayant pas d'équivalents dans le système universitaire anglo-américain, tendent mécaniquement à devenir invisibles dans les calculs automatisés des outils bibliométriques proposés par le WoS, s'ils ne sont pas placés en début de ligne d'affiliation.

Dans la configuration actuelle, 3 niveaux d'incertitude pèsent sur la chaîne de production des articles : [chercheur / chartes institutionnelles / chaîne éditoriale] ; un triangle dont les interférences aggravent la problématique des affiliations.

Bilan des solutions

La construction d'indicateurs fiables et représentatifs de la production scientifique française nécessite une nette amélioration du dispositif[12]. Des efforts considérables ont été faits et les diverses actions déjà entreprises sont riches d'enseignements. Les solutions envisagées sont, chacune, emblématiques d'une façon particulière d'analyser le problème.

Solutions fondées sur le facteur humain

- **Les chartes**. Leurs mesures incitatives sont suivies avec modération par les chercheurs, surtout lorsqu'elles sont agencées en séquences combinatoires. Le chercheur

[9] UMR : Unité Mixte de Recherche ; UPR : Unité Propre de Recherche
[10] EPST : Établissement à caractère scientifique et technique
[11] EPIC : Établissement public à caractère industriel et commercial
[12] Indicateurs de la recherche et politique documentaire : les documentalistes en première ligne [Dossier], *Documentaliste-Sciences de l'information*, novembre 2009, n°4.

et/ou son équipe peuvent être confrontés à plusieurs configurations possibles sans pour autant savoir choisir laquelle appliquer. Si sensibiliser (voire responsabiliser) les chercheurs doit rester l'une des composantes essentielles de la gestion des affiliations dans le paysage scientifique français, cette solution n'est pas le seul levier pour augmenter la visibilité de la production scientifique de la France au niveau international ni pour accroître l'attractivité des organismes de recherche et d'enseignement supérieur.

- **Un répertoire national**. Parfois imprécis et lacunaire parce que récent, le Répertoire National des Structures de Recherche (RNSR [13]) peut résoudre à terme une partie significative du problème. Mais là encore, la collecte des informations repose sur des interventions manuelles d'origines diverses.

- **Validation par les organismes**. Afin d'établir les indicateurs de la LOLF, l'OST fait valider les données du WoS par les organismes. Ce repérage, effectué par une personne connaissant parfaitement l'organisation de l'entité, garantit la qualité des données utilisées pour construire les indicateurs de la LOLF.

Solutions techniques

Le problème, posé sous un autre angle, peut mener à la prise en compte des technologies du Web de données (Linked Data), du Web sémantique et de l'interopérabilité des données.

- **Le projet national Conditor**[14] vise à recenser l'ensemble de la production scientifique française de la communauté « Enseignement Supérieur et Recherche ». Outre les données du WoS et des archives ouvertes HAL, la compilation obtenue pour l'année 2011 comporte plusieurs réservoirs institutionnels (CNRS, IRD, INRA, INRIA, Université Dauphine, etc.). Un traitement sophistiqué lui permet d'aligner des métadonnées entre ces différentes sources. La méthode d'appariement entre notices des différents corpus est effectuée automatiquement, sans validation manuelle du résultat. Des référentiels partagés (IdRef, structures et adresses CNRS, RNSR) enrichissent l'identification. L'intervention humaine a lieu principalement en amont, lors de la construction des outils bibliographiques et des référentiels.
Cette utilisation experte des métadonnées est également visible dans les logiciels du CWTS de l'université de Leiden (créatrice du classement éponyme) qui exploite de manière intensive la base WoS sans recourir à son champ *Organization-Enhanced*. Les algorithmes de Leiden ont été utilisés pour la mise au point du classement U-Multirank[15]. Mais, dans ce cas aussi, des vérifications manuelles restent incontournables.

- **Les mesures palliatives des éditeurs de bases de données**. Elles restent insuffisantes. Thomson-Reuters ajoute dans la base WoS un champ « intermédiaire » : l'*Organization-Enhanced* qui fait référence à toutes les graphies d'un seul et même organisme. On a donc quatre champs indexés sur le WoS : *Address* (AD) (sous-entendu du/des laboratoire(s) d'affiliation des auteurs), *Organization* (OO), *Suborganization* (SG), et *Organization-*

[13] RNSR https://appliweb.dgri.education.fr/rnsr/index.jsp?INIT=OK
[14] Projet conduit de janvier 2013 à avril 2014 initié par le Ministère de l'enseignement supérieur et de la recherche, dans le cadre de la BSN3, Bibliothèque Scientifique Numérique
http://www.bibliothequescientifiquenumerique.fr/
[15] http://www.umultirank.org/?trackType=home#

Enhanced (OG). Cependant quatre interrogations avec le terme CNRS engendrent quatre réponses différentes qu'il faut savoir «savamment» interpréter.
Interrogation du WoS 16-10-2014 avec la syntaxe «*Indexes=SCI-EXPANDED, CPCI-S, CPCI-SSH Timespan=All years»*

Champs interrogés (cf ci-dessus)	Nombre de publications
og=cnrs Organization-Enhanced [Index]	174,372
sg=cnrs Suborganization	342,620
ad=cnrs Address	659,952
oo=cnrs Organization	195,603

Une analyse plus fine du champ *Organization-Enhanced* démontre que les relations entre structures, sous-structures de recherche et variantes respectives peuvent manquer de rigueur. Les organismes voulant être correctement référencés doivent s'adresser à Thomson-Reuters pour valider et mettre à jour le contenu du champ *Organization-Enhanced*. Par ailleurs, les performances de l'outil d'analyse bibliométrique «InCites», module complémentaire du WoS, impliquent que ce champ soit correctement renseigné, ce qui s'avère être, dans les organismes, une manœuvre chronophage.

L'éditeur Elsevier a annoncé la création dans Scopus d'un traitement automatisé des affiliations, avec la promesse d'un gain de temps inégalé à la clé. Un test de cette interface révèle qu'un même organisme peut avoir en réalité plusieurs identifiants et qu'un contrôle d'unicité est nécessaire. Par ailleurs, l'interface s'en tient à l'interrogation stricte des établissements. Les laboratoires, instituts et départements ne font pas systématiquement partie des affiliations interrogeables.

Propositions, recommandations et perspectives

Ces solutions intéressantes sont loin de régler le problème des affiliations dans sa globalité. Sous certains aspects, comme on l'a vu, le remède peut être pire que le mal et engendrer de nouvelles contraintes, voire des facteurs de confusion, dans l'écriture des affiliations. Abordons alors le problème sous l'angle des technologies de l'information. Techniquement trois avancées semblent indispensables voire indissociables :

- Proposer aux éditeurs de bases de données la mise en œuvre d'**un module d'analyse** où les organismes ou établissements de recherche et d'enseignement seraient reconnus quel que soit leur positionnement dans les lignes d'adresses des publications et sans manipulation supplémentaire. Rechercher un terme dans un champ précis dans une base de données, en l'occurrence identifier les différentes composantes des adresses dans les champs *Address* ou *Organization* est une opération informatique triviale : une simple fonction d'extraction de chaînes de caractères délimitées par des séparateurs de texte. On reconnaîtrait ainsi qu'une adresse peut être articulée autour de plusieurs organisations ou tutelles principales, ce qui n'est pas pris en compte actuellement dans la fonctionnalité *Analyze* du WoS[16]. Il serait plus logique et simple que les spécifications techniques des éditeurs commerciaux s'adaptent aux exigences de l'organisation de la recherche et de leurs "utilisateurs-clients" plutôt que le contraire. Avec cette solution, la problématique mono ou

[16] Cf. recommandation RT 8 du rapport *Normadresses* OST mentionné plus haut ; page 23 de ce rapport

multi-ligne n'a plus lieu d'être, de même que la notion de première place dans une ligne d'adresse. On obtiendrait ainsi une uniformisation rapide et consensuelle au niveau national voire international.

- Demander aux éditeurs une **solution logicielle** permettant *a posteriori* le regroupement des affiliations par similarité ou alignement sémantique à partir de leurs bases de données. Il existe des progiciels suffisamment sophistiqués dont la souplesse du paramétrage s'adapte relativement bien à la problématique des affiliations.

- Utiliser **des identifiants uniques, pérennes** sous lesquels toutes les tutelles des structures mixtes ou collectives seraient déclarées permettant l'interopérabilité des différents systèmes. Ainsi, l'ensemble des unités de recherche cosignataires d'une publication seraient prises en compte sans avoir à les décliner de façon exhaustive et à rallonger démesurément les adresses. Les instances françaises se chargeraient de transmettre officiellement et régulièrement ces données structurées et à jour aux éditeurs de bases de données pour constituer et enrichir des tables d'affiliations intègres répondant à la réalité du système français.
Certains identifiants montent en puissance sur la scène internationale, comme l'ISNI[17] qui identifie les acteurs de la Recherche. À l'instar du RNSR, l'ISNI a quelques inconvénients majeurs : le manque de garantie en termes de qualité et de fiabilité des informations et l'absence de filiations entre les unités.
L'avantage de ces référentiels et identifiants réside dans les formulations alternatives. Leur inconvénient principal est le manque de renvois aux anciennes ou futures appellations, ainsi qu'aux partenaires, comme dans la base ISSN pour les titres de périodiques. Le numéro regroupant les partenariats avec une date de début et une date de fin serait à envisager.

Les solutions techniques, à portée de main, doivent s'accompagner de recommandations cohérentes et de mesures de sensibilisation auprès de chercheurs et des instances décisionnaires en insistant sur les enjeux de la démarche. Parmi celles-ci, on privilégiera :

* une écriture courte et stable car l'écriture des affiliations se heurte à des contraintes de longueur. L'utilisation du répertoire des structures RNSR faciliterait l'harmonisation de l'écriture des noms des laboratoires et de leurs tutelles. Mentionner, dans les adresses des auteurs, uniquement le nom du laboratoire et de ses tutelles et non les structures de regroupement géopolitiques ou financières apparaissant dans les remerciements et informations administratives à la fin de l'article.

* une information harmonisée et concertée vers les chercheurs et enseignants-chercheurs des laboratoires qui éviterait le choix cornélien de la "bonne" écriture des adresses.

Conclusion

Si la façon de rédiger et de retranscrire une affiliation a des conséquences majeures sur la production d'indicateurs, l'optimisation essentiellement technique des outils et le développement de référentiels restent incontournables. Il conviendra de jouer sur la complémentarité des solutions et l'interopérabilité des outils entre éditeurs et acteurs de la recherche.

[17] International Standard Name Identifier défini par la norme ISO 27729:2012.

Même si l'utilisation des identifiants se développe, le danger est de voir se multiplier ces référentiels sans liens entre eux, comme pour les chartes de signature. L'interopérabilité constitue une solution à mettre en œuvre, à l'instar de la proposition de HAL relative aux identifiants chercheurs dans sa version 3, car elle constitue une approche constructive pour établir des liens entre ces référentiels.

Michèle Dassa (a)
michele.dassa@cnrs-dir.fr
Alina Deniau (a)
alina.deniau@gmail.com
Manuel Durand-Barthez
Urfist de Paris
manuel.durand-barthez@enc.sorbonne.fr
Françoise Girard (a)
francoise.girard@polytechnique.edu
Nathalie Pothier (a)
nathalie.pothier@cnrs-orleans.fr
Angèle Sene (a)
angele.sene@cea.fr

(a) Réseau Renatis, Mission pour l'Interdisciplinarité, CNRS.