ASSIST: Outil pour l'extraction et l'analyse statistique d'articles

Justine Fouillé, Thi Lan Huong Nguyen, Baptiste Alix, Brett Becker, Matthieu Rochard, Hélène de Ribaupierre, Laurent d'Orazio

ASSIST: Outil pour l'extraction et l'analyse statistique d'articles

Justine Fouillé[*,**], Thi Lan Huong Nguyen[*,**], Baptiste Alix[*,**], Brett Becker[*,**], Matthieu Rochard[*,**], Hélène de Ribaupierre[*], Laurent d'Orazio[**]

[*]Cardiff University, Wales
deRibaupierreH@cardiff.ac.uk,
[**]Univ Rennes, CNRS, IRISA, France
laurent.dorazio@univ-rennes1.fr

Résumé. Il y a moins d'auteures que d'auteurs dans la recherche scientifique. Cependant, il n'existe pas encore, à notre connaissance, de système permettant différentes analyses sur les données disponibles et de confirmer cette hypothese. Ce travail propose une extension d'un outil précédemment réalisé, afin de le rendre plus performant et d'ajouter de nouvelles fonctionnalités. Ces nouvelles fonctionnalités sont le nuage de mots-clés ou la nouvelle fonctionnalité statistique. Les sources, références et autres informations sur l'article seront affichées pour chaque article récupéré. Les sexes des auteurs seront déterminés à l'aide d'une base de données reliant les prénoms aux sexes, afin de pouvoir obtenir des statistiques sur un grand nombre d'articles collectés.

1 Introduction

Publier ou périr illustre assez bien la pression que ressentent les chercheurs pour réussir une carrière universitaire. Les chercheurs sont ainsi encouragés à promouvoir leurs progrès en rédigeant des articles en les soumettant à des ateliers, des symposiums, des conférences ou des revues, au niveau national ou international.

Des études ont récemment démontré qu'il y a moins d'auteures que d'auteurs dans le domaine de la recherche scientifique [1]. Cependant, il n'existe pas encore de système fournissant des données concrètes pour le prouver, permettant par exemple de connaître la proportion de femmes pour un domaine donné (par exemple la médecine ou l'informatique), un sous-domaine (les bases de données ou l'interaction humaine par exemple), un domaine de recherche (comme l'intégration de données ou l'optimisation des requêtes), ou selon les différents évènements et selon leur classement (A*, A, B, C ou non classé). Il est donc impossible de voir l'évolution de la place des femmes au cours du temps.

1. Dans cet article, nous ne considérons que les femmes et les hommes, et des recherches supplémentaires devraient être menées pour inclure tout le monde.

Plusieurs plateformes existent pour analyser certains aspects des publications scientifiques telles que Scopus [2], altmetric [3], Plumx [4], Dimensions [5]. Ces plateformes portent soit sur les performances des scientifiques (comme le h-index), soit sur les réseaux de citations des articles, soit sur le classement de la revue (comme les facteurs d'impact). Des solutions telles que Core [6] regroupent les articles en libre accès de différentes ressources et proposent une API pour accéder au texte brut des articles. Cependant, à notre connaissance, il n'existe pas de plateforme ouverte permettant de rassembler et d'analyser les données contenues dans les articles scientifiques. Un cas d'utilisation possible de cette plateforme que nous présentons est d'analyser l'évolution de la place des femmes au sein de la communauté scientifique.

Dans cet article, nous présentons notre application Web, AXIS, pour Article eXtraction and statIstical AnalysiS. L'objectif est de dresser un portrait de l'évolution du monde de la recherche et de mettre en évidence les différences de genre dans ce domaine. AXIS permet donc de centraliser plusieurs sources bibliographiques et d'analyser les métadonnées des articles scientifiques. Les données concernant les articles (titre, liste des auteurs, références, citations, éditeur, origine...) sont donc récupérées depuis de multiples Interfaces de Programmation d'Applications ou API (DBLP, Core, etc.) pour varier les sources d'informations. Le sexe des auteurs est déterminé par leur prénom. L'application permet ensuite d'établir des statistiques sur le nombre d'auteurs par genre ou encore par évènement.

La structure de cet article est la suivante. Dans un premier temps, nous présentons comment nous améliorons les analyses faites avec les articles. Nous avons développé de nouvelles fonctionnalités permettant des analyses plus précises comme le nombre d'auteures selon les années chez un éditeur. Nous avons également développé un nuage de mots-clés pour chaque article qui affiche les principaux thèmes qu'il aborde. Pour finir, nous montrons comment pour résoudre le problème de récupération de données, nous avons changé le système de récupération des données des API.

2 Motivation

Les données bibliographiques extraites d'articles scientifiques ont plusieurs finalités. Les auteurs dans Di Iorio et al. (2015); West et al. (2013) proposent une liste des différentes tâches possibles sur les bibliographies selon les différentes classes d'utilisateurs. Dans cet article, nous aborderons deux tâches (construire une bibliographie et analyser l'impact d'un chercheur) et développerons la dernière en une nouvelle tâche (analyser l'évolution du paysage des publications et le nombre de femmes). Ce qui suit n'est pas une liste exhaustive de tous les outils et recherches possibles qui sont disponibles pour accomplir ces tâches, mais ceux qui sont souvent utilisés par les scientifiques pour accomplir ces tâches.

L'utilisateur souhaite constituer une bibliographie sur un sujet donné. Par exemple, l'utilisateur souhaite constituer une bibliographie sur le thème des "nanopublications" et obtenir toutes les publications pertinentes sur ce sujet. Des outils tels que Google Scholar [7] pourraient

2. https ://www.scopus.com/home.uri
3. https ://www.altmetric.com
4. https ://plumanalytics.com/
5. https ://www.dimensions.ai
6. https ://core.ac.uk
7. scholar.google.com

les aider à démarrer leur tâche. Les nanopublications étant un terme très spécifique, le résultat de la requête dans google Scholar ne sera que de 973 résultats. Si le scientifique veut explorer l'ensemble complet des résultats, cela lui prendra déjà un certain temps. Cependant, une étude montre que les scientifiques n'explorent souvent pas tous les résultats Renear et Palmer (2009). La tâche devient déjà plus délicate pour le scientifique puisque le premier article présenté à l'utilisateur est un article qui compte 286 citations. L'utilisateur souhaite-t-il explorer toutes les citations de cet article, ou seulement un sous-ensemble de celles-ci ? Et ce n'est que pour la première publication présentée au scientifique. Google Scholar offre une certaine possibilité de filtrage, comme la recherche parmi les 286 citations du premier article, mais n'offre pas la possibilité de filtrer à un niveau plus profond ou de rechercher tous les articles qui auraient pu être cités couramment par l'ensemble des 973 des articles. L'utilisateur commencera à étudier article par article, et regardera les différentes listes de citations, ouvrira les articles qui lui semblent intéressants, et cherchera les références ou les citations de ces articles. La tâche devient très longue et difficile car le nombre d'articles croît de manière exponentielle. Le scientifique pourrait alors changer pour un outil qui pourrait être plus approprié, et proposer plus de filtrage tel que Microsoft Academic [8]. L'interface vous permet de suivre un peu plus longtemps le fil des citations, mais si le scientifique souhaite télécharger un lot d'articles connexes, ce n'est pas possible non plus. De plus, il n'est pas possible de donner le nombre d'itérations que le scientifique souhaite réaliser dans sa recherche. A noter, que l'écosystème est dynamique. Ainsi Microsoft Academdic a été désactivé en décembre 2021.

La deuxième tâche consiste à analyser l'impact d'un scientifique. Cela se fait souvent par le nombre de citations sur les publications de l'auteur, ou de nouvelles métriques comme la popularité de cet article sur les réseaux sociaux (Tweeter, ResearchGate, etc). Plusieurs outils peuvent aider à accomplir cette tâche tels que Web Of Science [9], Google Scholar, Microsoft Academic, PlumX [10], Altmetric [11], Sci2Tool [12]. La plupart de ces outils ne donnent qu'une vue partielle de l'impact d'un scientifique et utiliseront différentes méthodes. Par exemple, Google Scholar semble compter dans le h-index du scientifique les auto-citations, quand d'autres outils comme WebOfScience ne comptent que les publications qui leur sont accessibles. Le chercheur qui veut analyser son propre impact ou l'impact de quelqu'un d'autre naviguera (lorsqu'il y aura accès) à travers différents sites Web et rassemblera les différentes informations pour pouvoir avoir une vue complète.

La troisième tâche consiste à analyser la participation des différentes minorités à la science. De nombreuses recherches montrent que le pourcentage d'auteures est inférieur à celui des hommes Danell et Hjerm (2013), cependant cette recherche est souvent effectuée de manière très manuelle, où le scientifique va collecter un sous-ensemble de données bibliographiques et analyser le pourcentage d'auteures de ce sous-ensemble. Cela prend beaucoup de temps et, en plus, ne montre qu'une vision très restreinte de la situation dans son ensemble. La comparaison entre les différents domaines de recherche est difficile, et pour pouvoir analyser à plus grande échelle, les données doivent être disponibles et traitées de manière automatique. Cette tâche présente plusieurs défis. Premièrement, toutes les données ne sont pas librement disponibles ; deuxièmement, toutes les données ne sont pas situées aux mêmes endroits ; et troisièmement,

8. academic.microsoft.com
9. https ://app.webofknowledge.com
10. https ://plumanalytics.com
11. https ://www.altmetric.com
12. https ://sci2.cns.iu.edu/user/index.php

certaines données doivent être déduites, telles que le sexe de l'auteur, car elles ne sont pas disponibles à partir des données de publication elles-mêmes.

3 Aperçu

Notre proposition, Article Extractions and Statistical Analysis (AXIS), se compose de deux fonctionnalités principales : (1) **Gestion de données** (Extraction et traitement de données) qui récupère les articles et répond aux requêtes d'articles pour les utilisateurs ; (2) **Analyse de données** permettra à l'utilisateur de collecter des statistiques sur les articles stockés et de les afficher.

4 Gestion de données

4.1 Extraction de données

La figure 1 illustre les flux de données au sein du système AXIS, d'un point de vue gestion de données.

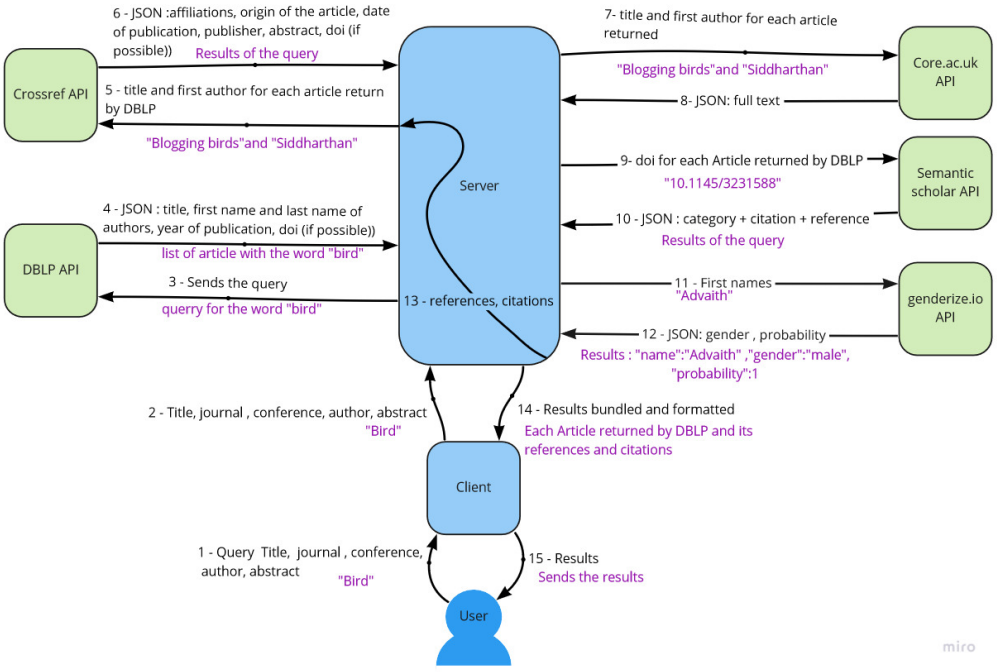

FIG. 1 – *Diagramme des flux de données*

Le système d'extraction de données est basé sur la combinaison de nombreuses API (telles que l'API DBLP, CrossRef et l'API Core). Ces API permettent d'accéder à des documents de

recherche provenant de différentes sources. En effet, chaque API ne fournissant pas les mêmes informations, ou fournissant des informations incomplètes sur les articles, combiner plusieurs API permet de récolter un maximum d'informations.

Pour chaque **article** son *titre*, *date de publication*, *résumé*, *DOI* [13] et le *texte complet* sont collectés. Pour les **auteurs** : leurs *nom*, *genre*, *position* (premier auteur, deuxième auteur...) et *affiliation* sont également stockés. Enfin, l'**origine** de l'article (une conférence ou une revue) ainsi que les *citations* de l'article et les *références* faites par l'article sont également collectées.

Une fois les articles récupérés, une application de définition du genre (telle que gender.io) permettra à la base de données d'associer un genre au nom de chaque auteur.

4.2 Difficultés

Les principales difficultés sont premièrement de **gérer les différents formats** des informations envoyées par ces API. Certains d'entre eux utilisent XML, d'autres JSON ou encore Bibtex. Une façon d'éviter ce problème consiste à n'utiliser que des API qui utilisent le même format ou peuvent renvoyer des formats différents. Une deuxième façon consiste à convertir les données dans un format unique une fois qu'elles sont reçues.

Le **manque de DOI** pour les articles plus anciens peut également poser un problème pour la récupération des citations et des références et donc des API alternatives peuvent être nécessaires. Les **homonymes** posent un autre problème car de nombreux auteurs pourraient partager le même nom et l'objectif sera de les distinguer. Une autre perspective sera de considérer le **changement de prénom ou de nom** au cours de la carrière de l'auteur.

Même si ce ne sera pas la réponse ultime à ces défis, nous pouvons utiliser l'identifiant ORCID d'un auteur. Actuellement, aucune des API qui ont été choisies ne renvoient l'identifiant ORCID de l'auteur, mais on peut imaginer que ce sera bientôt le cas, étant donné que son utilisation tend à augmenter.

5 Analyse de données

Après avoir recueilli beaucoup de données, il est possible de les analyser et de découvrir des informations intéressantes. En croisant une ou plusieurs sources de données, il sera possible d'obtenir des statistiques variées.

5.1 Visualisation

En utilisant les données stockées dans la base de données, des statistiques peuvent être calculées de différentes dimensions : le sexe de l'auteur, la position de l'auteur sur l'article, le nombre de citations, le nombre d'auto-citations, le nombre de références, les éditeurs, l'origine de l'article et la date de publication. Cette rubrique permet d'analyser différentes problématiques concernant le champ de la recherche scientifique et notamment la place des femmes dans celui-ci. L'utilisateur peut afficher le nombre d'auteurs par sexe par année pour un éditeur ou une conférence préalablement choisi. Le résultat est affiché sous forme de tableau avec le nombre total d'articles, le nombre d'auteures, le nombre d'auteurs et le nombre d'auteurs dont

13. Un identifiant d'objet numérique est un identifiant persistant ou un descripteur utilisé pour identifier des objets unique, normalisé par l'Organisation internationale de normalisation

le sexe est inconnu. Cela permet de récupérer des données brutes à utiliser dans diverses analyses. Grâce à cette fonctionnalité, il est possible de comparer le nombre d'auteurs par genre d'une année sur l'autre ou de plusieurs éditeurs une année donnée.

5.2 Nuage de mots clés

Le but du nuage de mots clés d'afficher les mots-clés d'un article scientifique mais aussi de récupérer ceux qui correspondent à ce mot-clé. Cela permet à l'utilisateur de visualiser les différents sujets qui correspondent à un terme spécifique. Cette fonctionnalité permet de découvrir de nouveaux articles scientifiques en proposant une liste d'articles correspondant au mot clé sélectionné par l'utilisateur. Ce nuage permettrait aussi de savoir sur quels sujets les femmes publient le plus. On aurait donc des statistiques encore plus précises que la proportion d'auteures dans un domaine.

5.3 Difficultés

Il y a plusieurs aspects qui peuvent influencer les statistiques, en particulier pour le sexe. Tout d'abord, il n'est pas toujours possible d'**attribuer un genre à un nom**. Pour cette raison, un seuil de probabilité de 85% de genderize.io a été décidé. Il semble préférable de construire des statistiques sur moins de données, plutôt que de construire sur de fausses données. Dans certains cas, il est également difficile de déterminer le sexe des auteurs, puisqu'ils ne fournissent que l'initiale du prénom.

6 Conclusion

Cet article présente AXIS, une plateforme s'appuyant sur plusieurs API pour rassembler et stocker de grandes quantités de publications scientifiques. Cette plateforme devrait aider le jeune chercheur à déterminer une bibliographie de manière beaucoup plus simple, car le processus récursif permettra à l'utilisateur de rassembler une liste précise et exhaustive de publications scientifiques.

Cette plateforme aidera également les scientifiques intéressés par l'analyse du paysage des publications scientifiques à avoir une vision plus complète ensemble de données à analyser.

De plus, la grande quantité de données permet l'analyse des statistiques et l'étude des tendances dans la communauté scientifique. Faciliter l'accès aux vastes quantités d'informations disponibles est primordial.

Avec les grandes quantités de données viennent cependant différentes contraintes telles que la contrainte des homonymes. Il est fort probable que parmi les auteurs insérés dans la base de données, certains porteront le même nom. Il n'existe actuellement aucun moyen de les distinguer. De plus, si un auteur change de nom, il sera enregistré comme deux personnes différentes. De plus, la plateforme n'analyse pour le moment que le sexe des auteurs en utilisant le prénom de l'auteur, cette manière d'analyser le sexe des auteurs conduira certainement à l'exclusion d'autres minorités, et devrait être investiguée plus avant.

Cette plateforme pourrait être très utile pour mener de nouvelles recherches dans le domaine de la scientométrie, ou simplement pour référencer un grand nombre d'articles scientifiques et leurs auteurs.

Références

Danell, R. et M. Hjerm (2013). Career prospects for female university researchers have not improved. *Scientometrics 94*(3), 999–1006.

Di Iorio, A., R. Giannella, F. Poggi, S. Peroni, et F. Vitali (2015). Exploring scholarly papers through citations. In *Proceedings of the 2015 ACM Symposium on Document Engineering*, pp. 107–116.

Renear, A. H. et C. L. Palmer (2009). Strategic reading, ontologies, and the future of scientific publishing. *Science 325*(5942), 828–832.

West, J. D., J. Jacquet, M. M. King, S. J. Correll, et C. T. Bergstrom (2013). The role of gender in scholarly authorship. *PLoS ONE 8*(7), e66212.

Summary

There are fewer female authors than male authors in the field of scientific research. However, there is not yet a system that provides a way to analyze the data that is available, and to backup that claim. This paper illustrates the upgrade of a tool previously made, in order to make it more efficient and add new features. Such new features are the keywords cloud or the new statistical functionality. Sources, references and other information on the article will be displayed for each articles retrieved. Genders of the authors will be determined using a database linking first names to genders, to be able to get accurate statistics on a large number of gathered articles.