AXIS - Plateforme d'Extraction d'Article et Analyses Statistiques

Brett Becker*, Matthieu Rochard*, Thi Lan Huong Nguyen*, Hélène de Ribaupierre**, Laurent d'Orazio*

*Univ Rennes, CNRS, IRISA, Lannion, France
laurent.dorazio@univ-rennes1.fr
http://www.iut-lannion.fr/
**Cardiff University, Cardiff CF10, Wales
deRibaupierreH@cardiff.ac.uk

Résumé. Une bibliographie est un élément fondamental dans un projet de recherche, en particulier pour définir un problème et motiver une proposition. Cet article de positionnement illustre le fonctionnement d'un outil permettant d'extraire des bibliographies. Il permet également de visualiser des résultats d'analyses statistiques sur des articles scientifiques. Cet outil propose un affichage récursif des références des articles et permet aux utilisateurs de parcourir toutes les sources de l'article. Le sexe des auteurs est déterminé automatiquement. Cela permettra de développer des statistiques précises sur un grand nombre d'articles.

1 Introduction

Effectuer des analyses sur les articles scientifiques est une tâche longue et complexe de par la grande quantité d'informations à collecter et à traiter. Les types de documents analysables sont très variés, il peut s'agir d'articles d'atelier, de colloques, de conférences, de revues, ou de rapports tels que des thèses de master ou de doctorat, sur un sujet écrit par un auteur donné. La quantité d'informations à traiter connaît également une croissance rapide. En effet, ces dernières années, le nombre de possibilités de soumettre des articles de recherche a fortement augmenté. à titre d'exemple, le nombre de documents de recherche référencés par DBLP était d'environ 25 000 en 1990, 76 000 en 2000, 223 000 en 2010, et était d'un peu moins de 379 000 en 2019. Dans un tel contexte, comment est-il possible d'aider les chercheurs à analyser une telle quantité d'informations, d'en savoir plus sur l'évolution de la recherche dans le monde et en particulier comment les femmes [1] chercheuses ont été prises en considération ? Le manque de diversité,qu'il soit genré ou ethnique, dans la recherche scientifique, est une problématique de plus de plus prise en compte et analysée. Une application permettant la simplification de la récolte de données et d'analyses statistiques de ces données permettrait de mettre ces ressources à la disposition d'un public plus large et d'étudier ces différents composants de façon plus aisée.

1. Dans cet article, nous prenons une approche binaire pour le genre

Plusieurs solutions existent pour analyser certains aspects des publications scientifiques comme par exemple : Scopus [2], altmetric [3] [4]Galligan et Dyas-Correia (2013). Ces solutions sont pour la plupart axées sur l'aspect des réseaux de citations, et fournissent des analyses sur le nombre de citations d'une publication, l'indice de notation d'un auteur (h-index) ou encore sur d'autres formes de citations comme le nombre de Tweets liés à un auteur ou à un article. D'autres solutions telles que : Core [5] se concentrent sur le regroupement d'articles scientifiques en libre accès.

Dans un premier temps, pour mesurer l'évolution de la présence des femmes dans la littérature scientifique, une collecte manuelle de données est nécessaire, et doit s'effectuer à partir de différentes plate-formes et interfaces de programmation d'applications (API). Dans un deuxième temps, une chercheuse [6] va devoir choisir quelles méthodes statistique sont les plus appropriées pour analyser ces données. Dans un troisième et dernier temps, cette chercheuse va choisir une technique de visualisation qui lui permettra, par exemple, de comprendre cette évolution au fil du temps et des domaines de recherche. à notre connaissance, une plate-forme fournissant ces différents services n'existe pas.

La solution imaginée est un système qui combinera de nombreuses API pour rassembler une quantité maximale d'informations pour chaque article, analysera les données et construira une visualisation efficace. Par cet intermédiaire, la base de données du système va regrouper toutes les informations disponibles sur un article, ce qui va rendre cette base comme l'une des plus complètes en terme de quantité d'informations sur les articles scientifiques. Le système est également en mesure de rassembler toutes les références et citations des articles, récursivement. Le sexe des auteurs est déterminé grâce à une API qui donne la possibilité de déterminer le sexe de l'auteure grâce à son prénom. Cela permet, par exemple, d'établir des statistiques sur le pourcentage de chaque sexe dans la communauté scientifique.

Cet article est structuré de la manière suivante : la première section traite de l'extraction et du traitement des données, en expliquant comment procéder pour permettre la collecte d'un maximum d'informations. En seconde partie, nous verrons comment analyser ces données, ce qui peut induire différents défis liés à l'identification des auteurs.

2 Motivations

Les données bibliographiques extraites d'articles scientifiques ont de nombreuses applications. Elles peuvent servir à l'élaboration d'une bibliographie d'un champ de recherches, ou à la méta-analyse de ce champ de recherches ou de la science dans son global. Dans cet article, nous allons prendre deux cas d'utilisations : l'analyse de la place des auteures femmes dans les publications scientifiques et son évolution, ainsi que le calcul de l'impact d'un chercheur sur sa communauté.

Pour calculer l'impact d'une publication scientifique, il faut souvent se baser sur le nombre de citations de celle-ci. D'autres métriques peuvent aussi être utilisées, telles que la popularité des articles sur les médias sociaux (Tweeter, researchGate, etc). Plusieurs outils existent

2. https ://www.scopus.com/home.uri
3. https ://www.altmetric.com
4. https ://plumanalytics.com/
5. https ://core.ac.uk
6. Dans cet article, nous utiliserons le féminin comme genre neutre.

déjà, tel que Web Of Science [7], Google Scholar, Microsoft Academic, PlumX [8], Altmetric [9], Sci2Tool [10]. La plupart de ces outils ne procure qu'une vue partielle de l'impact d'un scientifique et utilise des méthodes différentes. Par exemple, Google scholar semble compter dans le h-index du scientifique les auto-citations, alors que d'autres outils tels que WebOfScience [11] ne comptent que les publications qui leur sont accessibles. Un chercheur voulant analyser son propre impact ou l'impact de quelqu'un d'autre devra parcourir (quand il aura accès possible) différents sites Web et devra rassembler les différentes informations afin d'obtenir une vue complète de son impact.

Le deuxième cas d'utilisation de cette application consiste à analyser la participation des différentes minorités dans la recherche scientifique. De nombreuses recherches montrent que le pourcentage de femmes auteures est inférieur à celui des hommes Danell et Hjerm (2013); de Ribaupierre (2020). Cependant, cette recherche est souvent réalisée manuellement. La scientifique va ainsi procéder pour collecter un sous-ensemble de données bibliographiques et analyser le pourcentage de femmes auteures dans ce sous-ensemble. Cette méthode est coûteuse en temps et, de plus, elle ne donne qu'une image réduite de la situation globale. La comparaison entre les différents domaines de recherche est alors difficile, et afin de permettre une analyse à plus grande échelle, les données doivent être disponibles et traitées de manière automatique. Ce travail présente plusieurs défis. En premier lieu, toutes les données ne sont pas disponibles gratuitement ; ensuite, toutes les données ne sont pas situées aux mêmes emplacements ; enfin, certaines données doivent être déduites, comme le sexe de l'auteure, car elles ne sont pas disponibles à partir des données de publication elles-mêmes.

3 AXIS

Le système réalisé : Article eXtraction and statIstical analysiS (AXIS), est décomposé en deux sous-systèmes : L'extraction et le traitement des données qui récupèrent les articles correspondant aux requêtes de l'utilisatrice. Le second sous-système permet d'analyser les données et de proposer des méthodes statistique sur les articles récupérés et de les visualiser.

Extraction des données et traitement. Le but du système (Image 1) est de permettre à une utilisatrice d'effectuer une recherche sur des articles scientifiques via une interface Web. Les articles sont récupérés depuis différentes bases de données de sites bibliographiques (comme celles de DBLP, CORE, Semantic Scholar et Crossref). Les données récupérées concernent aussi bien les articles, leur type (journaux ou conférences) que leurs auteures. Les informations fournies par les API sont parfois similaires ; cependant, certaines API fournissent des informations plus précises. Par exemple, Semantic Scholar présente des informations plus détaillées sur les références et citations des articles. Dans ce cas précis, le système utilise en priorité les données renvoyées par Semantic Scholar concernant les références. Ces informations, une fois récupérées, sont regroupées puis stockées dans notre base de données. Les articles stockés sont alors consultables sur l'interface Web.

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

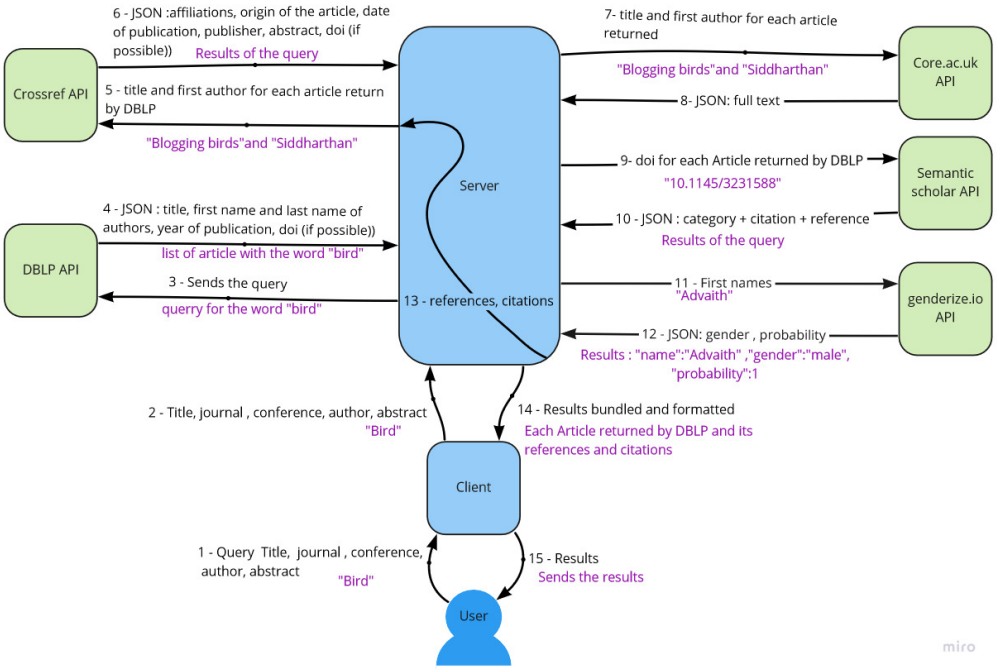

FIG. 1 – *Flux de données*

Processus. L'utilisatrice effectue une recherche à partir de l'application sur le titre, le journal, le nom de la conférence, le nom de l'auteure ou le résumé d'un article. La requête est envoyée à l'API de DBLP qui retourne une liste d'articles correspondant à la requête. Les données renvoyées sont les suivantes : le titre, l'année de publication et les noms des auteures de chaque article ainsi que le DOI lorsque celui-ci est disponible. Chaque article est ensuite vérifié par rapport aux articles déjà stockés dans la base de données de l'application, en utilisant le titre et le nom des auteurs. Si un article ou des éléments de celui-ci manquent dans la base de données, alors les autres API CrossRef, Core et Semantic Scholar sont appelées. Les données récupérées sont ajoutées aux données connues de l'article. Certaines API sont cependant privilégiées pour certaines données. CrossRef fournit les affiliations des auteurs, le résumé, une date de publication exacte, un type, l'éditeur et le DOI. CORE fournit le texte de l'article. Semantic Scholar fournit les données sur les références et citations (quand elles ne sont pas proposées par CrossRef) ainsi que la catégorie de l'article. Les méta-données des références et des citations de l'article sont extraites et stockées (titre, auteure, année de publication). Chaque article référencé ou cité passe par un processus similaire à celui que nous venons de décrire, ainsi que leurs références et citations, et cela, de manière récursive. Lorsqu'une nouvelle auteure est ajoutée à la base de données, son prénom est envoyé à l'API genderize.io pour déterminer le sexe. Pour finir, une liste contenant toutes ces données récoltées est renvoyée à l'utilisatrice afin qu'elle soit consultée.

Exemple : recherche sur le mot "Birds". L'utilisatrice effectue une recherche sur les articles contenant le mot "birds" dans le titre. Lorsque la recherche est envoyée à DBLP, l'API retourne une liste d'articles correspondant à la recherche. Supposons que le premier article de cette liste (intitulé "Blogging Birds", premier auteur "Advaith Siddhartan", publié en 2019 et DOI numéro "10.11.45/3231566") ne figure pas dans notre base de données. Le système enverra alors "Blogging Birds" et "Siddharthan" à l'API de CrossRef afin d'obtenir l'affiliation de l'auteure de l'article, le journal dans lequel celui-ci a été publié, l'éditeur, et la date exacte de publication. "Blogging Birds" et "Siddharthan" seront ensuite envoyés à l'API de CORE pour récupérer le texte intégral de l'article Blogging Birds. Le DOI sera envoyé à l'API de Semantic Scholar pour récupérer les citations et références de l'article. Le prénom de chaque auteure de l'article sera par la suite envoyé à l'API Genderize.io afin de déterminer le sexe. Si "Advaith" est envoyé à l'API, le résultat renvoyé sera "mâle". Enfin, toutes les références et citations de l'article passeront par le même processus, et cela, récursivement depuis la deuxième étape (CrossRef). Toutes les informations récupérées seront alors enregistrées dans une base de données locale.

Défis. La première difficulté a été de manipuler les différents formats utilisés par les API. Certaines utilisent du XML, d'autres du JSON ou encore Bibtex. Une deuxième difficulté est que les articles ne comportent pas tous un numéro DOI, ce qui peut poser problème lors de la récupération des citations et références des articles.
Une troisième difficulté est liée aux API elles-mêmes. Le temps de réponse des API utilisées, certains pouvant être vraiment longs. Ainsi, la récupération des articles sera longue dans un premier temps, mais plus le nombre d'articles stockés dans la base de données sera important, plus le temps de recherche diminuera pour les utilisateurs. Pour l'instant afin de limité le temps de recherche, le système ne récupère que les références récursivement à un premier niveau. Des outils telle que le parallélisme pourrait rendre possible la récurrence à un plus grand niveau. Cette dépendance aux API peut aussi poser problème lorsque celles-ci sont indisponibles, rendant parfois impossible la récupération de données. Une parallélisation du système est envisageable afin de récupérer les données sur plusieurs API en même temps, réduisant ainsi grandement le temps de récupération des données.
Une quatrième tient à l'identification unique des auteures. Les homonymes, le changement de nom possible au cours de leur carrière (mariage par exemple) peuvent être source d'erreurs. Nous pourrions utiliser l'ORCID id des chercheuses pour résoudre certains de ces problèmes d'identification ; cependant, un grand nombre d'articles, ou même d'auteurs, ne l'utilise pas.
Une cinquième difficulté est liée au fait que même en agrégeant les différentes API, il manque encore un certain nombre de données utiles. Par exemple, des articles provenant d'une même conférence sur des années différentes et dont le nom a changé selon les années, vont être classifiés différemment, la difficulté étant de regrouper ces informations. Pour surmonter cette difficulté, nous proposons une approche consistant à extraire les données directement des pages web disponible (web scraping ou web data extraction). Il est évident que ce processus peut prendre du temps, en particulier lorsqu'un grand nombre de pages est à traiter, ce qui entraîne une latence plus importante. Pour éviter une telle situation, nous proposons de stocker localement les données extraites. Les données collectées peuvent être importées dans la base de données locale pour une gestion efficace. Pour que ces données restent actualisées, des mises à

jour peuvent être envisagées périodiquement. Là encore, la récupération de données en parallèle améliorera grandement la vitesse de collecte. Pour le moment, nous extrayons de CORE le nom de la conférence, son acronyme, son code de catégorie et le nom de sa catégorie.

Analyse des données. Après avoir recueilli de nombreuses données, il est possible de les analyser et d'extraire des informations complémentaires en croisant une ou plusieurs de ces données. Les tests préliminaires ont été réalisés sur un nombre restreint d'articles, et sur un nombre restreint de méthodes statistique.

Processus. Les statistiques sont calculées sur les données stockées dans la base de données [12]. Les utilisatrices peuvent sélectionner les variables qu'elles souhaitent utiliser : tel que le sexe des auteures, la position des auteures dans l'article, le nombre de citations, d'autocitations et de références. Un graphique sera généré et l'utilisatrice pourra sélectionner les données qu'elles souhaitent voir apparaître sur chaque axe. Les utilisatrices peuvent choisir d'obtenir les données d'un axe sous forme de pourcentage ; dans ce cas, elles doivent également choisir une population pour le pourcentage.
Enfin, l'utilisatrice doit choisir le type de diagramme qu'elle souhaite afficher. La génération du graphique s'effectue à l'aide de l'API Google Charts. L'API utilisée pour l'affichage des diagrammes peut être changée étant donné que le système repose sur une architecture MVC [13].

Exemple : "Proportion de femmes par éditeurs" L'utilisatrice sélectionne "Proportion of Female Authors", "Editor" comme paramètre à visualiser et choisit un affichage en histogramme. La requête de la base de données retourne alors un jeu de données pour chaque paramètre. En croisant les deux jeux de données, on obtient "la proportion de femmes auteures par éditeurs". Ces données sont envoyées à l'API Google Charts et le graphique souhaité est renvoyé à l'application (Image 2) .

Défis. Il y a plusieurs aspects qui peuvent influencer les analyses, en particulier pour le genre. Tout d'abord, il n'est pas toujours possible d'attribuer un genre à un nom. Pour cette raison, un seuil de 85% de certitude du genre d'un nom a été fixé. Il semble préférable de réaliser une analyse sur moins de données, plutÃ´t que sur de fausses données. Dans certains cas, il est également difficile de déterminer le genre des auteurs, lorsque seules les initiales sont fournies, ou que le prénom est un prénom mixte tel que Charlie ou Claude.
L'analyse étant réalisée à partir des données contenues dans la base de données de l'application, elle est biaisée par le centre d'intérêt de l'utilisatrice et donc les recherches qu'elle a déjà réalisées avec AXIS. étant donné que des articles sont ajoutés à la base par les utilisatrices, cela nécessitera de nombreuses recherches afin de disposer d'un lot de données intéressant à analyser.

12. La base de données repose sur le système PostgreSQL
13. Nous utilisons le FrameWork PHP Laravel, respectant une architecture MVC

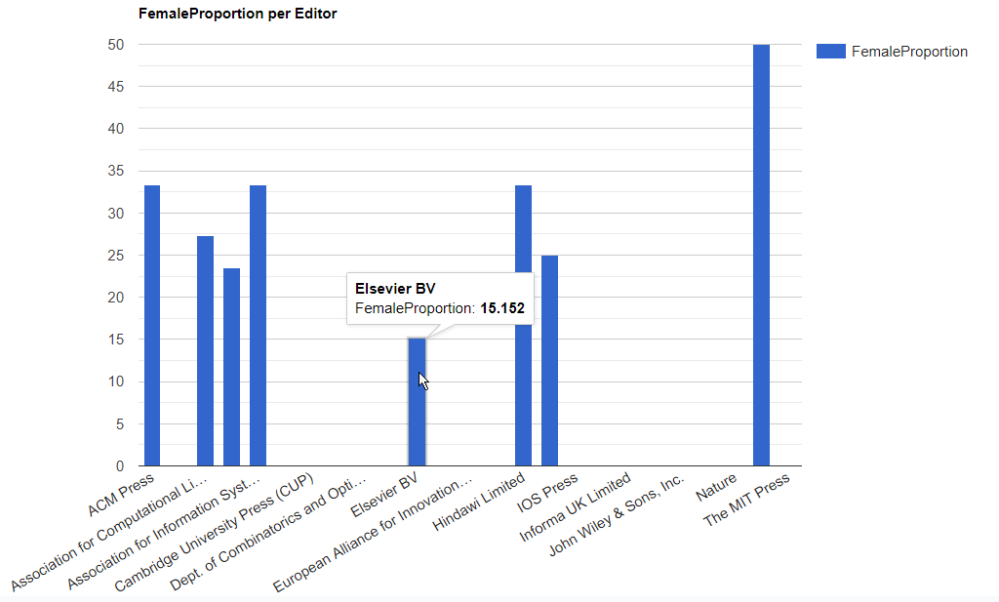

FIG. 2 – *Exemple de diagramme obtenu après l'analyse des données*

4 Conclusion

Cet article présente AXIS, une plate-forme reposant sur plusieurs API pour rassembler et stocker de grandes quantités de publications scientifiques. Cette plate-forme pourrait être très utile pour mener de nouvelles recherches dans le domaine de la sociométrie, ou simplement pour référencer un grand nombre d'articles scientifiques et leurs auteures. Le processus récursif devrait permettre à l'utilisatrice de rassembler une liste précise et importante de publications. Cette plate-forme aidera également les scientifiques intéressés par l'analyse du paysage des publications à avoir un ensemble de données à analyser.

La récolte automatique d'une importante quantité de données permet d'effectuer des analyses statistiques poussées sur des articles, et d'étudier les tendances dans la communauté scientifique. Il est primordial de faciliter l'accès aux grandes quantités d'informations disponibles. Cependant, la quantité des données à traiter s'accompagne d'un grand nombre de défis. Une des difficultés que nous ne sommes pas parvenus à résoudre est le problème des homonymes. Il est fort probable que parmi les auteures insérées dans la base de données, plusieurs porteront les mêmes noms. Il n'existe actuellement aucun moyen de les distinguer. De plus, si une auteure change de nom, elle sera enregistrée comme deux personnes différentes. Le développement de ce système nous a permis de prendre conscience des difficultés de réalisation des analyses sur les publications scientifiques. Ce projet a été également un enseignement sociétal, nous permettant de mieux nous représenter la place qu'occupent les femmes dans le monde de la recherche. Cet outil, en continuant de se développer, pourra mener à des études sur le genre, et sur l'évolution de la mixité dans le monde scientifique.

Références

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

de Ribaupierre, H. (2020). Egc, une confÃ©rence qui supporte la diversitÃ© genrÃ©e ? *Revue des Nouvelles Technologies de l'Information Extraction et Gestion des Connaissances , RNTI-E-36*, 381–388.

Galligan, F. et S. Dyas-Correia (2013). Altmetrics : Rethinking the way we measure. *Serials Review 39*(1), 56–61.

Summary

A bibliography is a fundamental aspect of any research, in particular to identify a problem and motivate a proposal. This position paper illustrates the workings of a tool made to extract bibliographies, and visualise insightful statistical analysis results. A recursive display of the article's references allows users to read the sources of the article. The authors gender is automatically determined. This will make gender orientated statistics possible on a large number of articles.