ASSIST: Article eXtraction and statIstical AnalysiS

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

ASSIST: Article eXtraction and statIstical AnalysiS

Justine Fouillé
Univ Rennes, CNRS, IRISA
Lannion, France
justine.fouille@etudiant.univ-rennes1.fr

Thi Lan Huong Nguyen
Univ Rennes, CNRS, IRISA
Lannion, France
thi-lan-huong.nguyen@etudiant.univ-rennes1.fr

Baptiste Alix
Univ Rennes, CNRS, IRISA
Lannion, France
baptiste.alixetudiant.@univ-rennes1.fr

Brett Becker
Univ Rennes, CNRS, IRISA
Lannion, France
brett.becker@etudiant.univ-rennes1.fr

Matthieu Rochard
Univ Rennes, CNRS, IRISA
Lannion, France
matthieu.rochard@etudiant.univ-rennes1.fr

Hélène de Ribaupierre
University of Cardiff
Cardiff, United Kingdom
deRibaupierreH@cardiff.ac.uk

Laurent d'Orazio
Univ Rennes, CNRS, IRISA
Lannion, France
laurent.dorazio@univ-rennes1.fr

***Abstract*—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.**

***Index Terms*—Data Extraction, Data Analytic, Statistical Analysis**

I. Introduction

Publish or perish illustrates quite well the pressure researchers feel to succeed an academic career. Researchers are thus encouraged to promote their progress writing papers submitting it to international workshops, symposia, conferences or journals.

Some studies have recently demonstrate that there are fewer female authors than male authors in the field of scientific research[1]. However, there is not yet a system that provides concrete data to prove this, making it possible for example to know the ratio of women for a given domain (for example medicine or computer science), a sub-domain (databases or computer human interaction for instance), a a research field (like data integration or query optimization), or according to the different venues and with respect to their rank (A*, A, B, C or unranked). It is thus impossible to see the evolution of the place of women during the time.

Several plateform exist to anaylse some aspects of scientific publications such as Scopus[2], altmetric[3], Plumx[4], Dimensions[5]. These plateforms are either focusing on the performances of the scientists (such as the h-index), the citations networks for the articles or the ranking of the journal (such as the impact factors). Solutions such as Core[6] aggregate the open access articles from different resources, and offer an API to access the raw text of the articles. However, to the best of our knowledge, there is no open platform that make it possible to gather and to analyse the data contained in the scientific papers. One possible use case of this platform we are presenting is to analyse the evolution of the place of women within the scientific community.

In this paper, we introduce our web application, AXIS, that stands for Article eXtraction and statIstical AnalysiS. The aim is to draw a portrait of the evolution of the research world and to highlight the differences of gender in this field. AXIS therefore makes it possible to centralize several bibliographic sources and to analyze data behind scientific articles. The data concerning the articles (title, list of authors, references, citations, publisher, origin...) are therefore retrieved from multiple Application Programming Interfaces or APIs (DBLP, Core, etc.) to vary the sources of information. The gender of the authors is determined by their first names. The application then allows statistics to be compiled on the number of authors by genre per publisher.

The structure of this article is as follows. First, we present how we improve the analyses made with the articles. We have developed new functionalities allowing more precise analyses such as the number of female authors according to the years with a publisher. We have also developed a keyword cloud for each article that displays the main themes it addresses. To finish, we show you how we solve the data retrieval problem we have changed the system for retrieving data from the API's.

[1]In this article, we are only considering female and male, and further research should be done to include everyone

[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

II. Motivation

Bibliographic data extracted from scientific articles have many purposes, the authors in [2], [4] propose a list of the different tasks on bibliographies possible depending on the different classes of users. In this paper, we will discuss two tasks (building a bibliography and analysing the impact of a researcher) and expand the last one into a new task (analysing the evolution of the publications landscape and the number of women). The following is not an exhaustive list of all the possible tools and research that are be available to achieve these tasks, but the ones that are often used by scientists to achieve these tasks.

The user wants to build a bibliography on a given topic. For example, the user wants to build a bibliography on the subject of "nanopublications" and get all the relevant publications on this subject. Tools such as Google Scholar[7] could help them to start their task. Nanopublications being a very specific term, the result of the query in google scholar will be only 973 results. If the scientist wants to explore the full set of results it will already take them some time. However, study shows that scientists often don't explore all the results [3]. The task becomes already more tricky for the scientist as the first paper presented to the user is a paper that has 286 citations. Does the user want to explore all the different citations from this paper, or only a subset of these 286 citations? And this is only for the first publication presented to the scientist. Google scholar offers some possibility of filtering, such as searching through the 286 citations of the first paper, but is not offering the possibility to filter on a deeper level or to seek for all the articles that could have been cited commonly by the set of 973 articles. The user will start to go article by article, and look at the different lists of citations, open the articles that seem interesting, and look for the references or the citations of these articles. The task becomes very time consuming and difficult as the number of papers grow in an exponential way. The scientist could then change for a tool that could be more appropriate, and offer more filtering such as Microsoft Academic[8]. The interface does allow you to follow the thread of citations a little bit longer, but if the scientist wants to download a batch of related articles, this is not feasible either. In addition, it is not possible to give the number of iterations the scientist wants to achieve in their research.

The second task is to analyse the impact of a scientist. This is often done by the number of citations on the publications of the author, or new metrics such as the popularity of this article on social media (Tweeter, researchGate, etc). Several tools can help to achieve this task such as Web Of Science[9], Google Scholar, Microsoft Academic, PlumX[10], Altmetric[11], Sci2Tool[12]. Most of these tools are giving only a partial view of the impact of a scientist, and will use different methods. For example, Google scholar seems to count in the h-index of the scientist the self-citations, when other tools such as WebOfScience is only counting the publications accessible to them. The researcher that wants to analyse their own impact or the impact of someone else will navigate (when they have access) through different web sites and gather the different information to be able to have a complete view.

The third task is to analyse the participation of the different minorities in science. Numerous research shows that the percentage of women authors is lower than men [1], however this research is often done in a very manual way, where the scientist will collect manually a subset of bibliographic data and analyse the percentage of women authors in this subset. This is very time costly, and in addition, it shows only a very small image of the big picture. The comparison between the different areas of research is difficult, and to be able to analyse on a larger scale, the data need to be available and to be processed in an automatic way. This task present several challenges. First, not all the data are freely available; Second, not all the data are located in the same places; and third, some data need to be inferred such as the gender of the author as they are not available from the publication data itself.

III. Overview

Our proposal, article Extractions and Statistical Analysis (AXIS), consists of two main functionalities: (1) **Data Management** (Data extraction and Processing) which recovers articles and responds to queries for articles for users; (2) **Data Analysis** will allow the user to gather statistics on the stored articles and display them.

IV. Data management

A. Data extraction

Figure 1 illustrates the data flow diagram of AXIS, from the data management perspective.

The data extraction system is based on the combination of numerous API's (such as the DBLP API, CrossRef and the Core API). These API's allow access to research papers from different sources. Indeed, since each API does not provide the same information, or provides incomplete information on the papers, combining several API's allows a maximum amount of information to be gathered.

For each **article** its *title*, *date of publication*, *abstract*, *DOI*[13] and the *full text* are collected. For the **authors**: their *names*, *genders*, *positions* (first author, second author...) and *affiliation* are also stored. Finally, the **origin** of the article (a conference or a journal) as well as *citations* of the article and *references* made by the article are also collected.

Once the articles are retrieved, a gender-defining application (such as gender.io) will allow the database to associate a gender to each author's name.

[7] scholar.google.com
[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

[13] A digital object identifier is a persistent identifier or handle used to identify objects uniquely, standardized by the International Organization for Standardization

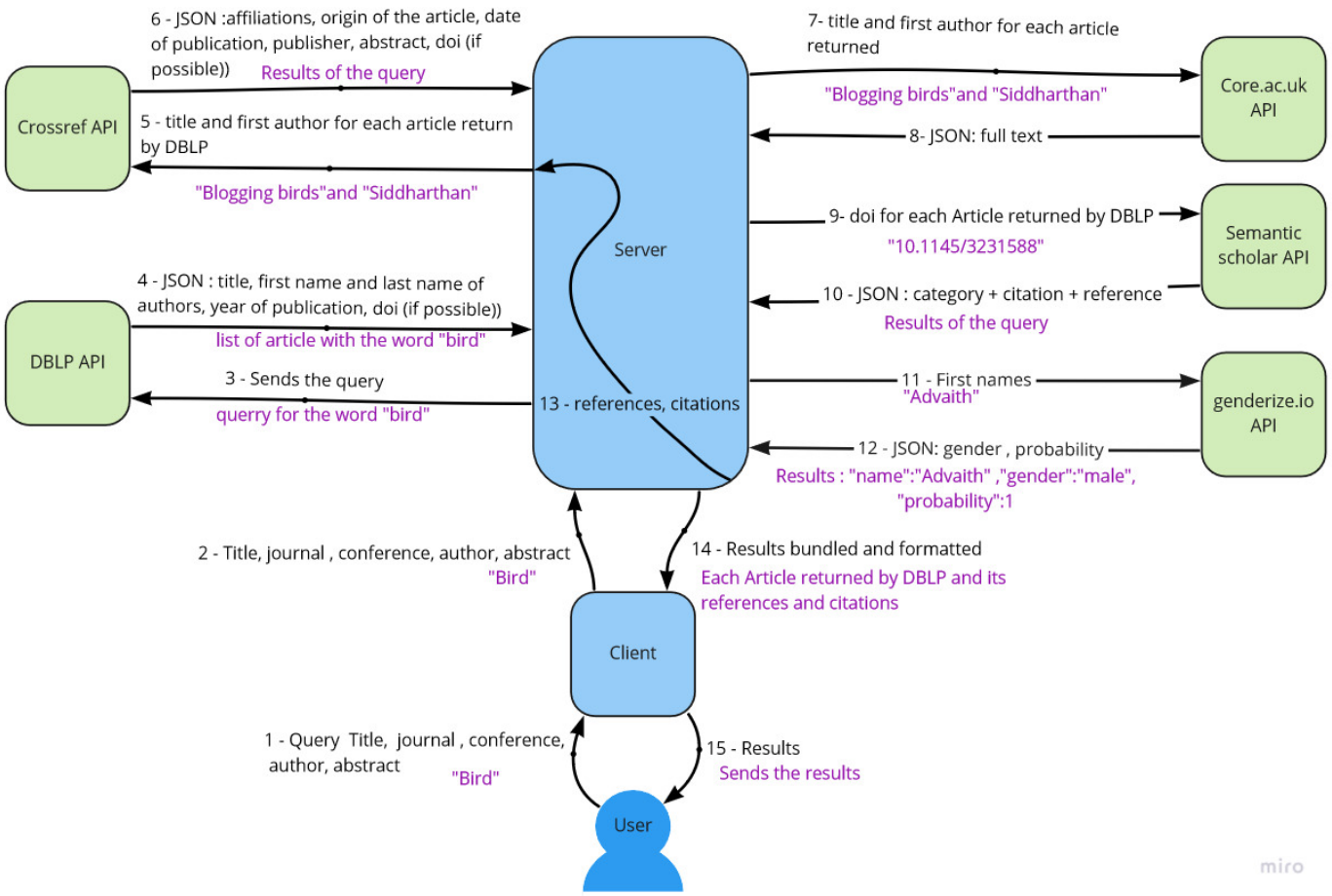

Fig. 1. Data flow diagram

B. Challenges

The main challenges are **managing the different formats** of information sent by these APIs. Some of them use XML, others JSON or even Bibtex. One way to avoid this problem is by using only APIs which use the same format or can return different formats. A second way is converting the data into a single format once it is received.

The **lack of DOI** for older articles can also pose a problem for recovering citations and references and thus alternative APIs may be required. **Homonyms** is another issue as many authors could share the same name and the challenge will be to distinguish between them. Another challenge will be the **change of first name or last name** during the career of the author.

Even if this will not be the ultimate answer to these challenges, we can use the ORCID id of an author. Currently, none of the API that have been chosen return the ORCID id of the author, but we can imagine that it will soon be the case, considering that its use is tending to increase.

V. Data Analytic

Having gathered much data it is possible to analyse it and uncover interesting information. By cross-referencing one or more data points it will be possible to obtain numerous statistics.

A. Visualization

Using the data stored in the database, statistics can be calculated from a variety of options: author's gender, author's position on the article, number of citations, the number of auto-citations, number of references, publishers, article's origin and the date of publication. This feature allows for the analysis of different issues concerning the field of scientific research and in particular the place of women in it. The user can display the number of authors by gender by year for a previously chosen publisher or conference. The result is displayed in the form of a table with the total number of papers, the number of men authors, the number of women authors and the number of authors whose gender is unknown. This makes it possible to retrieve raw data for use in various analyses. Thanks to this feature, it is possible to compare the number of authors per genre from one year to the next or from several publishers in a given year.

B. Keywords cloud

The tag cloud's purpose is to display the keywords of a scientific article but also to retrieve those that match that

keyword. This allows the user to view the different topics that match a specific term. This feature allows you to discover new scientific articles by offering a list of articles corresponding to the keyword selected by the user. This tag cloud would also allow to know on which topics women publish the most. We would therefore have even more accurate statistics than the proportion of female authors in a field.

C. Challenges

There are several aspects that can influence the statistics, in particular for the gender. First of all, it is not always possible to **attribute a gender to a name**. For this reason, a threshold of 85% probability from genderize.io has been decided upon. It seems preferable to build statistics on less data, rather than building on false ones. In some cases, it is also difficult to determine the gender of the authors, since they only provide the initial of the first name.

VI. Conclusion

This paper presents AXIS, a platform relying on multiple APIs to gather and store large amounts of scientific publications. This platform should help young researcher to determine a bibliography in a much easier way, as the recursive process will allow the user to gather a precise and exhaustive list of scientific publications.

This platform will also help any scientists interested in analysing the scientific publications landscape to have a more comprehensive data set to analyse.

In addition the large amount of data allows for the analysis of Statistics and the studying of trends in the scientific community. Facilitating access to the vast quantities of information available is paramount.

With the vast amounts of data however comes different constraints such as the constraint of homonyms. It is most likely that among the authors inserted into the database some will have the same names. There is no current way to distinguish between them. Moreover, should an author change names, they will be recorded as two different people. In addition, the plateform is currently only analysing the sex of the authors using the first name of the author, this way of analysing the sex of the authors will certainly lead to the exclusion of other minorities, and should be investigate further.

This platform could be very useful for conducting new research in the field of scientometry, or simply for referencing a large number of scientific articles and their authors.

References

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

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

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

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