

АВТОМАТИЗИРАНО ИЗВЛИЧАНЕ И АНАЛИЗ НА БИБЛИОМЕТРИЧНИ ДАННИ ЧРЕЗ API

Олга Маринова¹, Петя Страшимирова²

¹ Икономически университет – Варна/Катедра „Информатика“, Варна, България, olga.marinova@ue-varna.bg

² Икономически университет – Варна/Катедра „Информатика“, Варна, България, p.emilova@ue-varna.bg

РЕЗЮМЕ

Извличането на библиометрични данни от водещи наукометрични бази данни като Scopus и Web of Science става все по-важно за ефективното проследяване и анализ на научната продукция. Данните могат да бъдат автоматизирано събирани и обработвани въз основа на различни параметри, което ще позволи да ги използваме по-ефективно за целите на бъдещи обобщения и най-вече да получим задълбочена и достоверна представа за съвременните научни тенденции. Цел на доклада е да изследва възможностите на приложните програмни интерфейси (API) за извличане и анализ на библиометрични данни, както и да се предложи методология за автоматизирано извличане на данни от световните научни бази чрез API. Открояват се различни техники и методи за анализ на библиометрични данни.

КЛЮЧОВИ ДУМИ: извличане на данни, библиометричен анализ, SCOPUS API, Web of Science API

AUTOMATED EXTRACTION AND BIBLIOMETRIC DATA ANALYSIS WITH API

Olga Marinova¹, Petya Strashimirova²

¹ University of Economics – Varna, Varna, Bulgaria, olga.marinova@ue-varna.bg

² University of Economics – Varna, Varna, Bulgaria, p.emilova@ue-varna.bg

ABSTRACT

Retrieval of bibliometric data from leading scientific metrics databases such as Scopus and Web of Science is becoming increasingly important for effective tracking and analysis of scientific output. Data can be automatically collected and processed based on various parameters, which will allow us to use them more efficiently for the purpose of future summaries and, most importantly, to gain a thorough and reliable insight into current scientific trends. The aim of this paper is to explore the possibilities of Application Programming Interfaces (APIs) to retrieve and analyse bibliometric data, and to propose a methodology for automated data extraction from global scientific databases using APIs. Different techniques and methods for bibliometric data analysis are highlighted.

KEYWORDS: data extraction, bibliometric analysis, SCOPUS API, Web of Science API

ВЪВЕДЕНИЕ

В днешния динамично развиващ се свят науката търпи бурно развитие, при което в световен мащаб броят на изследванията и публикациите в различни платформи и реферирани списания нараства с изключителни темпове. Един от често използваните критерии за количествено и качествено измерване на научно-изследователската дейност са библиометричните

данни. С нарастващото количество публикувани материали, ръчният процес на събиране и анализ на данни става все по-трудоемък и неефективен. Тук на помощ идва автоматизацията чрез използване на автономен приложно програмен интерфейс (**Application Programming Interface – API**), който предлага бърз и структуриран достъп до световно известните научни бази от данни като Scopus и Web of Science. API позволява автоматизирано извличане на ключови библиометрични показатели като цитирания, автори, заглавия и ключови думи. Това значително улеснява анализа, който се провежда периодично, дори понякога ежегодно в академичните среди.

В тази връзка цел на доклада е да изследва възможностите на приложните програмни интерфейси за извличане и анализ на библиометрични данни, както и да предложи методология за автоматизирано извличане на данни от световните научни бази чрез API. Разглеждат се предимствата и функциите на API за извличане на библиографични данни с оглед имплементирането им в предлаганата методология за ефективно събиране и обработка на данни от двете (споменати по-горе) водещи световни научни бази.

Изследователската работа в доклада се провежда като част от проект „Усъвършенстване на процесите по дигитализация на научния ресурс посредством световните изследователски бази от данни“ (ДИГД) с № ПНИ-KC24-01-EDPSRGDB.

1. СЪЩНОСТ НА ИЗВЛИЧАНЕТО НА ДАННИ ЧРЕЗ API И ИЗПОЛЗВАНЕТО ИМ ЗА ПРЕДСТОЯЩИ АНАЛИЗИ

Библиометричните данни обикновено включват информация като брой на публикациите, данни за цитиранията, импакт факторите на списанията и други показатели, които отразяват научното влияние и значимост на даден изследовател или институция. Библиометричните анализи играят решаваща роля в разкриването на истинското въздействие на получените научни резултати по отношение на популяризирането и развитието на науката, технологиите, обществото и икономиката. Цитиранията и библиометричните показатели са основни компоненти в рейтинговите системи на университетите. Те не само предоставят обективна мярка за научната продукция, но също така са от ключово значение за репутацията, финансирането и стратегическото позициониране на университетите в световен мащаб.

Традиционните методи за събиране и анализ на множеството различни показатели, свързани с научната продукция е дълъг и бавен процес, който много често е съпроводен с допускането на грешки, дублирания или затруднения при намиране на нужната информация. От друга страна автоматизираните подходи за извличане на данни предлагат значителни предимства.

Ето защо темата за автоматизирано извличане и анализ на библиометрични данни чрез API се превръща в изключително актуална и значима в контекста на нарастващия обем научна литература и необходимостта от ефективни инструменти за управление на съпътстващия публикации поток от разнородна информация. Най-често автоматизираното извличане на библиометрични данни се изпълнява чрез използването на Application Programming Interfaces (API) или на уеб скрапинг (**web scraping**). Приложните програмни интерфейси (API) осигуряват стандартизиран интерфейс, който позволява на различни приложения да комуникират и да се свързват лесно, като рационализират процеса на разработване на софтуер и го правят по-ефективен.

Съществуват три основни възможности за достъп до API: безплатен, академичен и платен достъп. Фокусът на настоящето изследване е върху API и неговото използване за достъп до

библиометрични данни чрез API на научните бази Scopus и Web of Science, тъй като те предлагат както платен, така и безплатен достъп до някои основни библиометрични данни, които да послужат за последващ анализ във външни системи.

API интерфейсите на Scopus и Web of Science дават възможност за автоматизирано извличане на данни за публикации, цитирания, автори и институции. Вместо потребителите да търсят и копират нужната им информация ръчно, API прави заявки директно към огромния обем от научни данни и връща необходимата информация в структуриран формат (обикновено JSON или XML). Така можем да извлечем данни в реално време и да получаваме актуализирани библиометрични показатели. Освен това API предоставя изключителна гъвкавост при създаването на заявки. Можем да дефинираме много точни параметри за търсене: търсене по ключови думи, автор, институция, тема, обхват от години, вид публикация и много други. Това позволява изключително прецизно сегментиране на данните, като можем да изведем точно тази информация, която ни интересува и след това да я анализираме.

Scopus е всеобхватна и многофункционална база от данни, собственост на Elsevier, която предоставя на изследователите богат набор от академична литература и данни за цитирания, както и редица функции за подпомагане на сътрудничеството, откриването и оценката на научните изследвания. Това е една от най-големите и най-широко използвани академични бази данни (Bharathi and Aditya Sai Srinivas, 2024, p. 31) с над 90 милиона публикации, включващи научни списания, книги и материали от конференции от цял свят (Scopus.com, 2023). Тя също така предоставя инструменти, които институциите да използват за анализ на резултатите и въздействието на научните изследвания, както и API за разработчиците за създаване на персонализирани приложения и инструменти.

Web of Science (WoS) е другата световноизвестна база от данни, която се е наложила като една от най-често използваните в научните среди (Bharathi and Aditya Sai Srinivas, 2024, p. 37). Понастоящем е собственост на Clarivate Analytics и обхваща повече от 22 000 списания, книги и материали от конференции и над 90 милиона документа (Clarivate.libguides.com, 2024). Един от видовете API, които WoS предлага е **Web of Science Starter API**. Той осигурява REST-базиран програмен достъп до документите и списанията в Web of Science с JSON формат на изхода и основни възможности за търсене и филтриране. Чрез Web of Science Starter API се предоставя достъп в реално време до библиографски метаданни, като DOI, автори, заглавие на източника и други метаданни, които са в основната база на Web of Science, като и от свързани с нея бази данни. Той позволява изграждане на директни връзки от външни системи към статии в WoS и извличане на данни за броя на цитиранията.

При все това, трябва да отбележим, че повечето комерсиални бази данни, в това число и Scopus и Web of Science обикновено предоставят достъп до своите API с разширени възможности за извличане на библиометрични данни само чрез платен или институционален акаунт, което може да доведе до това, че изследователи или университети с по-ниски доходи да не могат да се възползват от всички тези данни. Друг съществен въпрос са редица правни аспекти, които следва да бъдат взети предвид във връзка с последващото използване на извлечените данни във външни системи, каквито са например университетските системи за отчитане на публикациите на академичния състав.

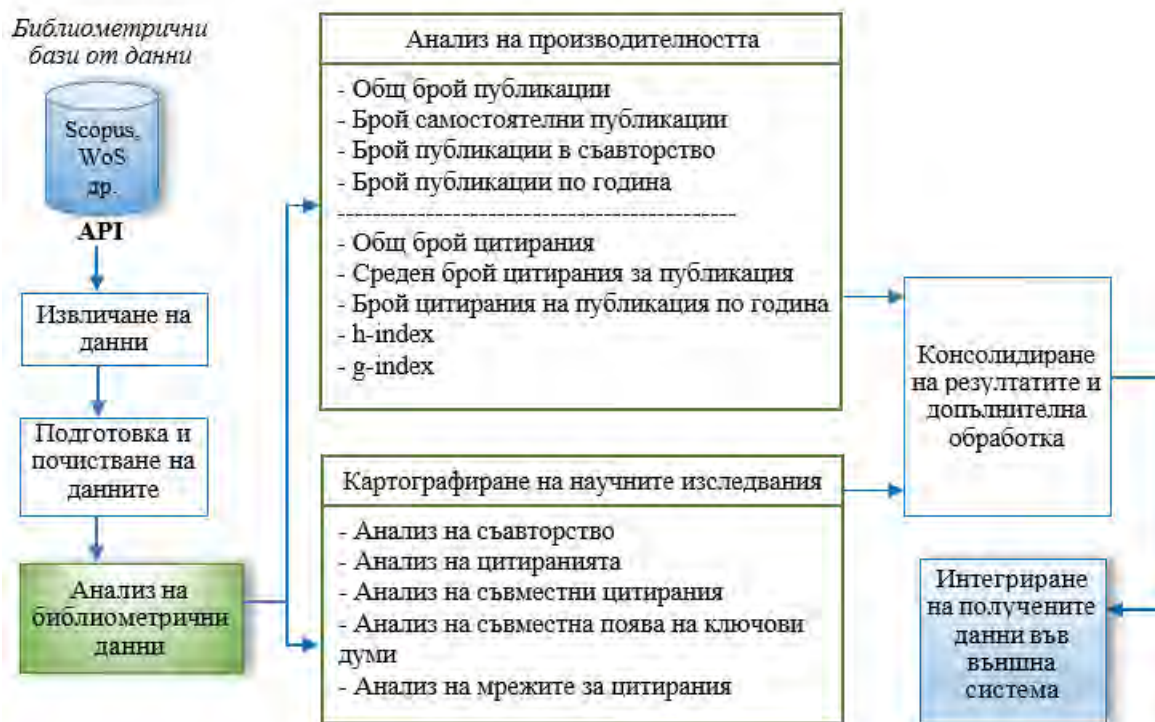
За да се изясни цялостния процес по извличане и използване на библиометрични данни е нужно да се очертаят всички стъпки, през които трябва да премине автоматизираният подход за събиране, обработка и анализ на данните.

2. МЕТОДОЛОГИЯ ЗА ИЗВЛИЧАНЕ НА ДАННИ

Автоматизираното извличане и последващ анализ на библиометрични данни позволява точен мониторинг на научната продукция, което е пряко свързано с измерването на глобалното въздействие и репутация на всеки един университет, както и в подкрепа на неговото движещо ядро – изследователите.

Въпреки че библиометричният анализ има много предимства, той е сравнително нов инструмент и все още недостатъчно добре обхванат в реномираните ръководства за систематични прегледи на литературата (Passas, 2024). В тази връзка считаме, че задълбочения преглед на различните видове анализи и създаването на подход, който да дава яснота относно данните, които биха могли да се събират чрез API на Web of Science и Scopus и последващата обработка върху тях биха били добра отправна точка за бъдещото разработване на система, която да извършва това автоматизирано извличане.

Оценяването на научното въздействие на изследователи и списания чрез редица индикатори като брой публикации, общ брой цитирания, брой цитирания на публикация по година е широко разпространена практика. В тази връзка още един показател е добил широка популярност, благодарение на своята лесна интерпретация – h-индексът, който е предложен от Hirsch (Hirsch, 2005). За да бъде един библиометричен анализ изчерпателен и всеобхватен той следва да обхване редица метрики и анализи, които можем да обособим в две основни категории – *Анализ на производителността (Performance analysis)* и *Картографиране на научните изследвания (Science mapping)* (Noyons et al., 1999). Освен тях в целия процес по извличане, анализ и интегриране на данни от библиометрични БД се включват още няколко важни стъпки. В тази връзка на фиг. 1 е предложена методология, която обхваща следните значими според нас процеса – автоматизираното извличане на данни чрез API; подготовка и почистване на данните; последващ библиометричен анализ на данните; консолидиране на резултатите и интегрирането им във външна система.



Фигура 1. Методология за извличане, анализ и интегриране на данни от библиометрични БД

Източник: Собствена разработка

Чрез ясно очертаване на отделните етапи се гарантира систематичност и възпроизводимост на резултатите, което повишава научната стойност на получените библиометрични показатели. При все това част от метриките, които се показани на фигура 1 не могат да бъдат извлечени директно чрез API на Scopus или WoS и изискват допълнителни изчисления и обработка на получените метаданни. Това са метриките **общ брой цитирания**, **среден брой цитирания за публикация** и **g-index**. Освен тях всички анализи, които са представени като част от групата **Картографиране на научните изследвания**, също могат да бъдат изпълнени само чрез изчисления с допълнителни програмни инструменти.

Можем да заключим, че всеки един елемент от библиометричния анализ спомага за изграждането на пълна и цялостна картина на даден изследователски или институционален профил и неговото въздействие в научната общност. Затова следва да отделим внимание на позадълбоченото вникване в същността на различните видове анализи на библиометрични данни.

3. АНАЛИЗ НА БИБЛИОМЕТРИЧНИ ДАННИ

Както беше разгледано по-горе библиометричният анализ включва редица метрики и анализи, които можем да разграничим в две основни категории: *анализ на производителността* и *картографиране на научните изследвания*.

3.1. Анализ на производителността

Оценката на резултатите в областта на научните изследвания включва оценка на въздействието на изследователите, институциите и държавите, като се използват показатели като общ брой публикации, принос на авторите и показатели, свързани с цитирането.

- **Общ брой публикации** – броят на публикации, издадени от даден изследовател, институция или държава.
- **Брой самостоятелни публикации** – публикации, написани от един автор, в които е посочен индивидуален принос в изследванията.
- **Брой публикации в съавторство** – публикациите, написани от няколко автори, които отразяват съвместни изследователски усилия.
- **Брой публикации по година** – годините, през които изследователят или институцията активно публикуват.
- **Общ брой цитирания** – общият брой цитати от публикациите на даден изследовател или институция.
- **Среден брой цитирания на публикация** – показател, който определя средния брой цитирания на една публикация и се използва за измерване на нейното въздействие в определена област или година. Този показател помага да се сравнява влиянието на публикациите в различни области или години чрез нормализиране на разликите в честотата на цитиранията.
- **Брой цитирания на публикация по година** – средният брой цитирания на всяка публикация за дадена година.
- **h-index** – h е стандартен научен показател, в който се съпоставят броят на публикуваните статии и броят на цитиранията на техния автор. Формулата се основава на броя на публикациите (H), които са цитирани поне H пъти. Той е лесен за разбиране и относително устойчив, тъй като не се влияе от прекалено високите цитирания на отделни публикации.
- **g-index** – индексът придава по-голяма тежест на високоцитираните статии. Той отчита

както броя на публикациите, така и броя на цитатите на една публикация. При даден набор от статии, подредени в низходящ ред според броя на получените от тях цитирания, **g**-индексът е (уникалното) най-голямо число, при което първите **g** статии са получили (заедно) поне **G²** цитирания (Egghe, 2006).

3.2. Картографиране на научните изследвания

Научното картографиране спомага за очертаването на структурата и динамиката на научните изследвания. То се осъществява чрез различни подходи като:

- **Анализ на съавторството (Co-Authorship Analysis)** – разглежда взаимодействията между учените в определена област, като изследва институциите и страните, които те представляват. С нарастващата сложност на изследванията сътрудничеството се превръща във все по-често срещана практика като същевременно допринася за по-ясни и задълбочени резултати. Чрез анализът на сътрудничеството се подпомага идентифицирането на групи от изследователи, които работят заедно или в сходни научни направления и които се свързват чрез общи съавтори и изследователски теми. По този начин може да се разкрие клъстеризация на изследвания в дадена област и да се насърчат нови изследвания в области, които не са достатъчно проучени. Също така проследява сътрудничествата във времето, което е полезно за бъдещи учени, търсещи партньорства с утвърдени или перспективни изследователи (Ponomariov and Boardman, 2016).
- **Анализ на цитиранията (Citation Analysis)** – ключов метод за картографиране на научната област, който измерва влиянието на публикация чрез броя на получените цитати, отразявайки интелектуалните връзки между публикациите. Той помага да се идентифицират най-влиятелните публикации и да се разбере интелектуалната динамика в дадена област, като цитатите остават най-обективният показател за въздействието на научната работа (Naveen et al., 2021). Ето защо чрез анализа на цитиранията могат да се изследват най-влиятелните публикации в дадена област и да се проследи интелектуалната динамика в нея.
- **Анализът на съвместното цитиране** – метод за научно картографиране, който разкрива тематичната свързаност между публикации, които често се цитират заедно. Той може да покаже интелектуалната структура на дадена област и да се идентифицират основните теми и идеи в нея. Предимствата на този метод са в разкриването на тематични клъстери – групи от публикации, които са тясно свързани по съдържание или изследователски фокус, както и да се открият най-влиятелните изследвания. От друга страна негово ограничение е, че често е насочен към по-старите, по-често цитирани публикации и по този начин по-новите или по-специфичните и непопулярни изследвания остават извън обсега му. Това го прави особено полезен за откриване на ключови знания в бизнес изследванията (Rossetto et al., 2018).
- **Анализ на съвместната поява на ключови думи** – техника за извличане на информация от текст, която изследва „съвместната поява“ на двойки ключови думи в преглежданите документи. Този подход предполага, че ключовите думи, които често се срещат заедно в едни и същи документи, имат определена връзка помежду си.
- **Анализ на мрежите за цитирания** – използва се за изследване на научната литература по дадена тема. С помощта на една публикация могат да бъдат открити допълнителни релевантни източници, които демонстрират както качествено, така и количествено връзките между статиите и техните автори.

4. ПРЕДИЗВИКАТЕЛСТВА И БЪДЕЩИ ТЕНДЕНЦИИ

Библиометричният анализ придобива все по-голямо значение в академичните среди, тъй като позволява обработката на огромни обеми научна литература и осигурява обективна мярка за влиянието на изследванията и публикациите. Чрез библиометричния анализ могат да бъдат идентифицирани основни изследователски тенденции, ключови теми и автори, които са популярни или добиват популярност в научните среди. Всичко това допринася за изграждането на по-ясна представа за структурата и развитието на научните изследвания.

Въпреки предимствата си, автоматизираното извличане на данни чрез API има своите технологични ограничения. При използване на API за достъп до бази от данни като Scopus и Web of Science често е необходима допълнителна обработка на данните, за да се извлекат и изчислят специфични метрики, като общ брой цитирания или g-индекс. Освен това, комерсиалните бази от данни обикновено предлагат платен достъп до разширените си API функции, което може да се окаже пречка за академичните институции с ограничени бюджети. Наличието на юридически аспекти, свързани с използването на данни от тези източници, също е фактор, който следва да се има предвид при разработване на нови методологии.

С напредъка и развитието на технологиите за обработка на големи масиви от данни, както и широкото използване на изкуствения интелект, библиометричният анализ придобива все по-голяма значимост като иновативен и перспективен подход за академичните изследвания. Комбинирането на новите технологии и възможностите на API могат да се използват за по-прецизно картографиране на научните области и открояване на ключови публикации и автори, както и за прогнозиране на научни тенденции и потенциални области за развитие. Автоматизацията в анализа ще продължи да улеснява научната общност, като предоставя ефективни и достъпни инструменти за мониторинг на научната продукция и въздействието ѝ върху обществото и икономиката.

ЗАКЛЮЧЕНИЕ

В заключение, автоматизираното извличане и анализ на библиометрични данни чрез API интерфейси предоставя ценни възможности за научната общност, като осигурява бърз и стандартизиран достъп до основни показатели като цитирания, брой публикации, h-индекс и други метрики. Всички тези метаданни чрез подходяща обработка и последващ анализ биха предоставили отлична основа за провеждането на допълнителни различни по сложност и детайлност анализи.

Ето защо настоящият доклад подчертава значението на автоматизираното извличане и анализ на библиометрични данни чрез приложни програмни интерфейси на водещи бази от данни като Scopus и Web of Science. Разработена е методология за ефективно събиране, обработка и анализ на данни, която цели да внесе яснота в цялостния процес по автоматизирано извличане на данни, анализиране и използване във външни системи. Подобна разработка следва да бъде в помощ на изследователите и институциите при измерването на научната продукция и нейното въздействие.

Включени са и подходи за анализ като картографиране на науката и анализ на съавторството, които разкриват структурните и тематични връзки в изследванията. Можем да обобщим, че с непрекъснатото увеличаване на броя на публикациите в световен мащаб автоматизацията на процесите по извличане на библиометрични данни придобива решаващо значение. Тя съкращава времето за ръчна обработка, минимизира риска от грешки и улеснява провеждането на по-задълбочен библиометричен анализ.

REFERENCES / ИСПОЛЗВАНА ЛИТЕРАТУРА

1. Bharathi, M. and Aditya Sai Srinivas, T. (2024) Scopus & WoS: The Ultimate Tool for Scholarly Research. *Advancement of Computer Technology and Its Applications*, 7(3), pp. 31-40. Available from: <https://doi.org/10.5281/zenodo.11195393/> [Accessed 12/10/2024].
2. Clarivate.libguides.com. Resources for Librarians and Administrators. Available from: <https://clarivate.libguides.com/librarianresources/coverage/> [Accessed 17/10/2024].
3. Egghe, L. (2006) Theory and practise of the g-index, *Scientometrics*, 69(1), pp. 131–152. Available from: <https://doi.org/10.1007/s11192-006-0144-7/> [Accessed 12/10/2024].
4. Hirsch, J. E. (2005) An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the U.S.A*, 102 (46), pp. 16569-16572. Available from: <https://doi.org/10.1073/pnas.0507655102/> [Accessed 12/10/2024].
5. Narong DK, Hallinger P. (2023) A Keyword Co-Occurrence Analysis of Research on Service Learning: Conceptual Foci and Emerging Research Trends. *Education Sciences*, 13(4) pp. 339. Available from: <https://doi.org/10.3390/educsci13040339> [Accessed 12/10/2024].
6. Naveen D., Satish K., Debmalya M., Nitesh P. and Weng Lim. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*. 133. pp.285-296
7. Noyons, E.C.M, Moed, H.F., Luwel, M. (1999) Combining mapping and citation analysis for evaluative bibliometric purposes: a bibliometric study. *J. Am. Soc. Inf. Sci.* 50(2), pp. 115–131. Available from: [https://doi.org/10.1002/\(SICI\)1097-4571\(1999\)50:2<115::AID-ASI3>3.0.CO;2-J](https://doi.org/10.1002/(SICI)1097-4571(1999)50:2<115::AID-ASI3>3.0.CO;2-J) [Accessed 15/10/2024]
8. Passas, I. (2024) Bibliometric Analysis: The Main Steps. *Encyclopedia 2024*, 4(2), pp. 1014-1025. Available from: <https://doi.org/10.3390/encyclopedia4020065/> [Accessed 16/10/2024].
9. Ponomariov, B. and Boardman, C. (2016). What is co-authorship? *Scientometrics*, 109, pp. 1939–1963 Available from: <https://doi.org/10.1007/s11192-016-2127-7> [Accessed 16/10/2024].
10. Rossetto, D.E., Bernardes, R.C., Borini, F.M., Gattaz, C.C. (2018) Structure and evolution of innovation research in the last 60 years: review and future trends in the field of business through the citations and co-citations analysis. *Scientometrics*, 115(3), pp. 1329 – 1363. Available from: <https://doi.org/10.1007/s11192-018-2709-7> [Accessed 15/10/2024].
11. Scopus.com. Scopus now includes 90 million + content records. Available from: <https://blog.scopus.com/posts/scopus-now-includes-90-million-content-records/> [Accessed 12/10/2024].
12. Villa-Collar, C., Alvarez-Peregrina, C., Martinez-Perez, C., & Sánchez-Tena, M. (2024) Citation network analysis on keratoconus. *Journal of Optometry*, 2(17), 100498. Available from: <https://doi.org/10.1016/j.optom.2023.100498> [Accessed 15/10/2024].