２Ｅ１１

Excel ベースの大規模書誌データ取得・分析支援ツールの開発

○山下泰弘，野澤龍介，花田文子，村野文菜（ＪＳＴ）
yasuhiro.yamashita@jst.go.jp

1. はじめに

今日では論文の引用索引データベースの高機能化が進んでおり、必ずしも書誌情報をダウンロードしなくても各種の分析が可能である。しかし、より詳細な分析を行うため、データをエクスポートしてデータセットを固定し、ローカル環境で精査と加工をするケースも多い。

エルゼビア社の Scopus は、Web インターフェースから 2 万件までの書誌情報をエクスポート可能としている。研究領域など大規模な文献集合を対象とした分析では、文献数が 2 万件を超えるケースが多く、このエクスポート件数の上限はボトルネックとなる。分割出力の機能はないので、2 万件超の書誌情報をダウンロードするには検索式を複数に分けるなどかなりの工夫が必要となる。また、同社の分析ツール SciVal は 20 万件の論文データを分析できるが、Scopus から SciVal に一度にエクスポートできる論文数の上限も 2 万件であり、それが SciVal での大規模データ分析作業を煩雑にしている。

そこで、より大規模な分析を容易にするため、Excel ベースで Scopus から API を介し大規模データを取得、分析するツール「bibliopad」を開発している。本稿ではその特徴と使用例を紹介する。

2. bibliopad - Scopus Search API Accessing Tool

2.1 Scopus Search API

bibliopad では、Web 版 Scopus の上限である 2 万件を超えたエクスポートを実現するため、Scopus Search API（高橋,2021; Elsevier, 2023）を使用する[1]。Scopus Search API は、Scopus ユーザの学術研究等のためにエルゼビア社が無償で提供している。1 クオータ当たり最大 200 件の論文データを取得可能で、ユーザには 1 週間当たり 20,000 クオータが割り当てられる。従って、ユーザは 1 週間に最大 400 万件の書誌データを抽出できる。同 API で返されるデータには、書誌タイトル、発表年、掲載誌、著者名、著者所属、著者キーワードなどが含まれる。著者名と著者所属が 1 論文当たり 100 名分までしか含まれない点は注意が必要である。また、学術分野（ASJC）と参考文献は含まれない。

2.2 bibliopad の概要

bibliopad は、分析者の技術的な障壁を下げることを企図している。そのため、大規模な書誌データを取得できることに加えて、(1) 追加ソフトウェアのインストールを要しないこと、(2) 特別な技能を必要としないことを要件とした。これらを達成するため、広く普及している Microsoft Excel をベースに、プログラミング言語 Visual Basic for Applications (VBA)で構築したマクロによりデータの取得、加工を行うこととした。さらに、通常 Excel のワークシート関数のみでは困難な共著関係等の集計機能を付加するとともに、熟練ユーザ向けには XML 形式の生データをダウンロードすることも可能にした。なお、計量書誌学分析ソフトウェアは多数公開されているが、Scopus からの書誌データ抽出と集計を Excel のみで完結しているものは、調査した限り見いだせなかった[2]。

bibliopad でのデータ取得および集計処理の流れは図 1 の通りである。まずユーザは検索式を記入し、Scopus に問い合わせを行う。この際に用いる検索式は Web 版 Scopus の詳細検索と同一なので、Web 版で基本検索を行い、詳細検索式を表示させたものをコピー・ペーストしてもよい。検索結果に著者が 100 名を超える論文が含まれる場合、アラートと、当該論文の検索式が表示されるので、必要に応じて Web 版で当該論文のデータを検索し、CSV 形式でエクスポートしたものを取り込み、著者名と国を補完できる（所属機関のデータは、Web 版で取得できるデータが完全には整理されていないため、補完できない）。

---

[1] API の利用に当たりエルゼビア社より個別に許諾を得た。

[2] 既存の分析ツールについては Moral-Muñoz, et al., (2020)を参照した他、Web 調査を行った。ジャーナル検索については VBAinExcel（https://github.com/haozeng0/VBAinExcel）が公開されているが、論文検索についてのツールは見いだされなかった（Web 調査は 2024 年 9 月 18 日実施）。

それを踏まえて、国・所属機関・著者といったアクターについて、論文数および共著論文数の集計を実行できる。

API で取得する書誌情報には著者および所属機関が含まれているが、完全に名寄せされているわけではない。そこで、著者 ID および機関 ID に基づき同一名称に名寄せする機能を付加した。Web 版からのエクスポートデータにも著者 ID は含まれるため、著者 100 名超の論文に関するデータ補完機能と併用もできる。

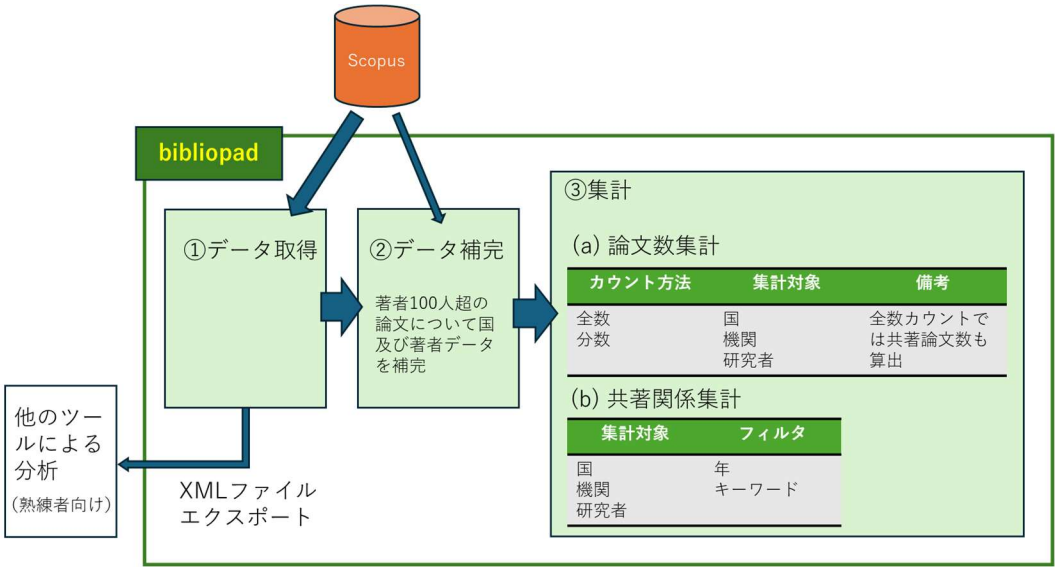


図 1　bibliopad の処理内容

2.3 bibliopad の検索画面

bibliopad の検索画面は、図 2 のようなものである。一番上に検索式を入力するセルがあり、その下に消費したクオータや、残りクオータが初期値にリセットされる日時等の情報が表示される（青いセルの項目）。ユーザは、これらの情報と、検索時に表示される必要クオータ数を考慮し、検索を実行するか中止するかを決定できる。

使用中、bibliopad に大量のデータが蓄積されて動作が遅くなった場合、「データの消去ボタン」で容易に初期状態に戻すことができるようにした。消去対象は、ダウンロードしたデータも含む全データ、bibliopad で算出した統計データのみ、の 2 通りを選択できる。

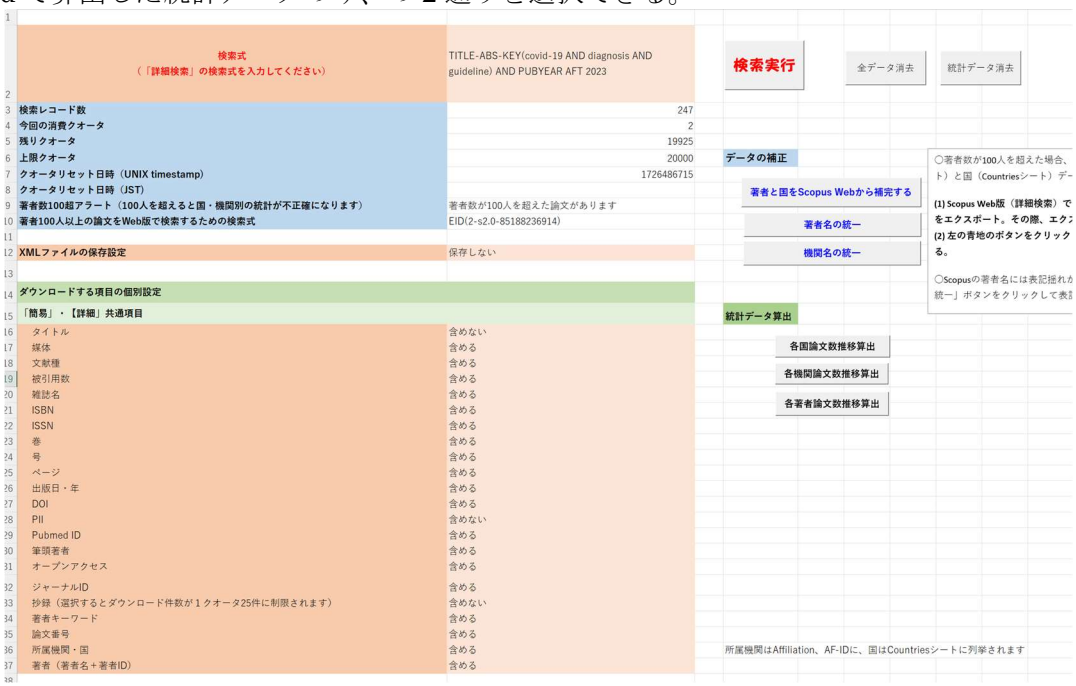


図 2　bibliopad の検索画面

2.4 集計機能

bibliopad には以下の集計機能を付加した。この機能は、分析者に Excel のみで作成することが困難な集計データを補完的に提供するものである。

(a) 論文数集計

論文数集計は、全数カウントと分数カウントの両方をサポートしている（図 3）。分数カウントは、論文中での出現回数でウェイト付けをした集計（例えば日本 1 機関、米国 2 機関が同じ論文に出現した場合、日本 1/3 件、米国 2/3 件とする。図 3 では「分数カウント（機関レベル）」と表記）、アクターの異なり数で按分した集計（先の例では日本 1/2 件、米国 1/2 件とする。図 3 では「分数カウント（国レベル）」と表記）の 2 種類を用意してあるが、年別の推移については前者についてのみ算出している。全数カウントについては、複数アクター間の共著論文数（図では国際共著）も算出される。

(b) 共著関係集計

Excel ユーザにとって、エクスポートしたデータの単純集計は容易であるが、複数アクター間の共著関係データを得ることは難しい。そこで、全数カウントでの論文数上位アクターについて共著マトリクスを作成する機能を実装した（図 4）。上から何位までを集計対象とするか、集計対象期間、キーワード（タイトル、抄録、著者キーワードに出現する文字列）で絞り込むことが可能である。キーワードによるフィルタリングをするためには、目的に応じてタイトル、抄録、著者キーワードのデータ項目をダウンロードしておく必要がある。

国	論文数（全数）	分数カウント（国レ	分数カウント（機関レベル）	国際共著論文
China	1978	1808.10119	1840.547597	308
United States	1009	719.1698773	712.9249662	503
Germany	278	161.0532107	155.3025781	190
United Kingdom	253	137.4198773	132.0175134	181
Japan	238	194.5	195.020011	77
Italy	212	138.7266234	143.0234294	119

図 3　論文数集計シート（左）

集計実行

集計方法	対角成分	国際共著論文
	含める国の論文数順位	10
集計期間	年（自）	記載が無い場合は西暦０年扱い
	年（至）	記載が無い場合は西暦30000年扱い
キーワード（タイトル、アブストラクト、著	RNA	記載が無い場合は全体

	China	United States	Germany	United Kir	Japan	Italy	India	Fran
China	17	7	1	0	1	0	0	
United States	7	27	4	2	2	4	2	
Germany	1	4	9	1	0	1	0	
United Kingdom	0	2	1	6	1	2	0	
Japan	1	2	0	1	2	0	0	
Italy	0	4	1	2	0	7	0	

図 4　共著関係分析シート（右）

2.5 bibliopad の使用感と制約

bibliopad で検索を行うと、1 クオータの処理に 4〜5 秒程度を要する。使用中は Excel のリソースを占有してしまうため、数万件以上の大規模データをダウンロードする際には、Excel を要する他の業務との兼ね合いに留意する必要がある。試験的に「aging（老化）」に関する 2001〜2025 年の論文 75,653 件のダウンロードと集計を行ったところ、ダウンロードに 379 クオータ、29 分を消費し、それらの文献の著者 284,395 名それぞれの論文数と推移の集計処理に 47 分を要した[3]。bibliopad により 20,000 件を超える論文データを取得できるが、Excel が扱うことができるデータは最大 1,048,576 行であるため、これが bibliopad で取得できる論文数の上限である。

3. bibliopad の使用例（アフリカ主要国の論文数分析）

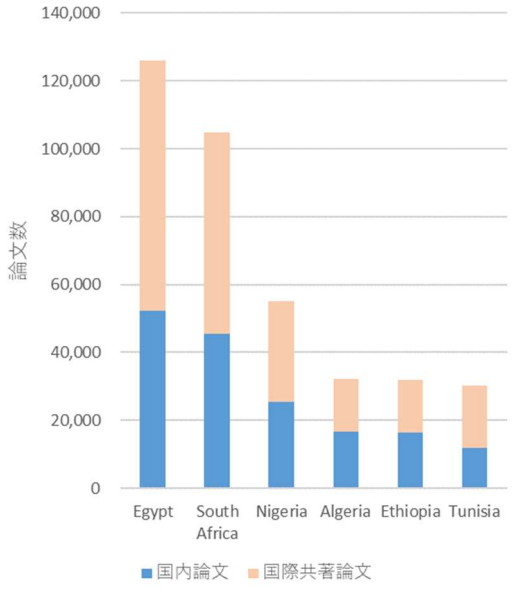
bibliopad を用いて、アフリカの論文数上位 6 カ国（エジプト、南アフリカ、ナイジェリア、アルジェリア、チュニジア）の 2021〜2023 年の 3 年間の論文等出版および国際共著の状況を分析した[4]。

総文献数は 367,264 件であり、全数カウントによる各国の文献数は図 5 の通りである。6 カ国中でエジプトと南アフリカの文献数が多く 10 万件超となっているが、その過半は国際共著によるものだった。一方、6 カ国中で 4 位のアルジェリア、5 位のエチオピアは過半が国内論文であり、国によって研究のありようが異なっていると考えられる。6 ヶ国の共著相手国を見ると、英国を旧宗主国とするエジプト、南アフリカ、ナイジェリア、フランスを旧宗主国とするアルジェリア、チュニジアがそれぞれ旧宗主国および旧宗主国を同じくする国との共著割合が高いことが分かる。一方、エジプトおよびチュニジアの論文に占めるサウジアラビアとの共著論文の割合はそれぞれ 28.1%、21.6%（全数カウント）と顕著に高

[3] bibliopad を用いて Scopus（https://api.elsevier.com および https://www.scopus.com）よりデータを取得（検索日 2024 年 9 月 19 日）。本稿で紹介するデータはすべて当該サイトより取得した。

[4] データ取得日は 2024 年 8 月 9 日。

く、同様に、エチオピアの論文に占めるインドとの共著論文の割合も 15.8%と高い。この要因については、科学技術を取り巻く国際動向についての背景理解とともに、書誌データの内訳などさらなる深掘り分析を進める必要があろう。



共著相手国	分析対象国					
	Egypt	South Africa	Nigeria	Algeria	Ethiopia	Tunisia
Egypt	–	1.1%	2.1%	3.5%	1.2%	3.9%
South Africa	0.9%	–	12.4%	1.0%	3.7%	2.0%
Nigeria	0.9%	6.5%	–	0.6%	2.1%	0.7%
Algeria	0.9%	0.3%	0.4%	–	0.3%	3.6%
Ethiopia	0.3%	1.1%	1.2%	0.3%	–	0.4%
Tunisia	0.9%	0.6%	0.4%	3.3%	0.4%	–
United States	8.6%	15.8%	10.7%	3.0%	9.6%	3.6%
Saudi Arabia	28.1%	2.3%	3.7%	8.3%	3.3%	21.6%
United Kingdom	5.0%	13.5%	9.8%	3.2%	6.0%	3.2%
India	5.4%	5.8%	6.3%	3.4%	15.8%	3.5%
Germany	3.7%	6.0%	2.9%	1.9%	3.6%	3.2%
China	7.3%	4.5%	5.2%	2.6%	3.8%	2.9%
France	1.8%	3.9%	1.5%	15.6%	1.5%	18.4%
Australia	1.9%	6.4%	2.9%	0.9%	3.9%	1.2%
Italy	2.5%	3.5%	2.0%	3.5%	1.7%	4.7%
Canada	2.4%	4.6%	3.0%	2.3%	2.2%	3.0%
Spain	1.8%	3.0%	1.2%	3.8%	1.2%	4.4%
Netherlands	0.8%	4.0%	1.3%	0.6%	2.1%	0.6%
Pakistan	4.4%	1.5%	2.6%	1.9%	1.6%	3.0%
Switzerland	0.8%	3.2%	1.6%	0.7%	1.7%	1.0%
Brazil	0.9%	2.4%	1.7%	0.7%	0.8%	1.1%
Sweden	0.8%	2.8%	0.9%	0.7%	1.8%	0.6%
Japan	2.6%	1.7%	1.2%	0.6%	1.0%	0.8%
Turkey	2.2%	1.3%	2.0%	3.8%	0.7%	2.2%
Malaysia	1.8%	1.3%	7.0%	1.5%	1.2%	1.2%
Russian Federation	1.9%	1.5%	1.0%	0.9%	0.5%	0.9%
Belgium	1.0%	2.4%	0.9%	1.4%	2.1%	1.3%
South Korea	1.8%	1.2%	1.0%	1.4%	1.6%	1.1%
United Arab Emirates	3.2%	1.1%	1.1%	2.2%	0.7%	2.2%
Iran	1.3%	1.6%	1.2%	1.6%	0.9%	1.6%

図 5　アフリカ 6 カ国の国内論文および国際共著論文の数（2021-2023）（左）
図 6　アフリカ 6 カ国の論文に占める各国との共著割合（分析対象国も含む上位 30 カ国）（右）

4. 今後の課題

bibliopad は今後も改良を進める予定である。現時点では以下が課題となっている。

(1) Web 版からのデータ補完機能などを追加したことにより、初版よりも使い方が複雑になった。ユーザフォームなどを活用し、より直観的なインターフェースにする必要がある。

(2) 現仕様では分野分類の付与を手動で行う必要があるが、分析しやすい形式でジャーナルリストから取り込めるようにすればより利便性が上がると考えられる。

(3) 現仕様では、著者別および機関別論文数集計において、Excel の行数上限超過を避けるため、著者や機関の異なり数が 1,000,000 件を超えた時点で集計自体を停止する設定となっているが、停止せずに論文数上位アクターのみをワークシートに表示することを検討している。

(4) 検索結果に極端に古い文献が含まれる場合、論文数推移集計のワークシートのサイズが非常に大きくなる。集計対象年をあらかじめ絞り込めるようにし、不必要なサイズの増加を抑制することも検討している。

本稿では使用例の一部を示したが、口頭発表時にはさらなる分析事例についても紹介したい。bibliopad の対外配布については、今後検討していく予定である。

参考文献

Elsevier. (2023, September), *Elsevier APIs getting started guide version1.* https://dev.elsevier.com/guides/Scopus%20API%20Guide_V1_20230907.pdf

Moral-Muñoz, J. A., Herrera-Viedma, E.; Santisteban-Espejo, A., & Cobo, M. J. (2020). Software tools for conducting bibliometric analysis in science: An up-to-date review. *El profesional de la informa-ción*, *29*(1), e290103. https://doi.org/10.3145/epi.2020.ene.03

高橋昭治. (2021). 出版から情報分析へ：Scopus がもたらした変化,情報の科学と技術. 71, 398-403. https://doi.org/10.18919/jkg.71.9_398