DISCUSSION PAPER No.162

Microsoft Academic Graph の書誌情報データとしての評価

Assessment of Microsoft Academic Graph as a Bibliographic Data Source

2018 年 10 月

文部科学省　科学技術・学術政策研究所
第 1 研究グループ
塚田尚稔・元橋一之

本 DISCUSSION PAPER は、所内での討論に用いるとともに、関係の方々からの御意見を頂くことを目的に作成したものである。

また、本 DISCUSSION PAPER の内容は、執筆者の見解に基づいてまとめられたものであり、必ずしも機関の公式の見解を示すものではないことに留意されたい。

The DISCUSSION PAPER series is published for discussion within the National Institute of Science and Technology Policy (NISTEP) as well as receiving comments from the community.

It should be noticed that the opinions in this DISCUSSION PAPER are the sole responsibility of the author(s) and do not necessarily reflect the official views of NISTEP.

【執筆者】

塚田 尚稔　　文部科学省科学技術・学術政策研究所 主任研究官

元橋 一之　　東京大学大学院工学系研究科 教授
文部科学省科学技術・学術政策研究所 客員研究官
独立行政法人経済産業研究所 ファカルティフェロー

【Authors】

Naotoshi Tsukada　　Senior Research Fellow, National Institute of Science and Technology Policy (NISTEP), MEXT

Kazuyuki Motohashi　　Professor, Graduate School of Engineering, The University of Tokyo
Affiliated Fellow, National Institute of Science and Technology Policy (NISTEP), MEXT
Faculty Fellow, Research Institute of Economy, Trade and Industry (RIETI)

本報告書の引用を行う際には、以下を参考に出典を明記願います。
Please specify reference as the following example when citing this paper.

塚田尚稔・元橋一之 (2018)「Microsoft Academic Graph の書誌情報データとしての評価」, *NISTEP DISCUSSION PAPER*, No.162, 文部科学省科学技術・学術政策研究所.
DOI: http://doi.org/10.15108/dp162

Naotoshi Tsukada and Kazuyuki Motohashi “Assessment of Microsoft Academic Graph as a Bibliographic Data Source,” *NISTEP DISCUSSION PAPER*, No.162, National Institute of Science and Technology Policy, Tokyo.
DOI: http://doi.org/10.15108/dp162

Microsoft Academic Graph の書誌情報データとしての評価

文部科学省　科学技術・学術政策研究所　第 1 研究グループ
塚田尚稔・元橋一之

要旨

本論文では Microsoft 社の書誌情報データ Microsoft Academic Graph（MAG）の利用可能性について、計量書誌学の分野で利用実績の多い Elsevier 社の Scopus をベンチマークとして大規模サンプルで評価した。Open Academic Society から無償ダウンロードできる MAG のバルクデータを用いて、各データベースの論文を DOI で接続して、同一論文 19,166,705 件の書誌情報を比較した。論文出版年は 97.0%、著者数は 98.8%の論文で一致した。参考文献数は Scopus の方が多いが、書誌情報がデータベースに収録されている参考文献に限ると MAG の方が多い。MAG と Scopus のそれぞれから求めた被引用数（2005 年出版の論文、出版後 10 年間の引用）は、スピアマン順位相関係数が 0.945 であり、高い相関を示す。一方、MAG の 1.66 億件の文献のうちで全ての著者に所属機関情報が存在する論文は 4,373 万（26.3%）であり、限定的である。MAG は全体として有用なデータベースであるが、現状では、所属機関情報を用いる研究などのためには商用データベースに頼る必要があると考えられる。

Assessment of Microsoft Academic Graph as a Bibliographic Data Source

First Theory-Oriented Research Group, National Institute of Science and Technology Policy (NISTEP), MEXT

ABSTRACT

We assessed Microsoft Academic Graph (MAG) as a bibliographic data source, comparing to Scopus of Elsevier as the benchmark. We used bulk data of MAG, which we can download for free from Open Academic Society. We matched documents in MAG and Scopus using DOI and compared the bibliographic data extracted from MAG and Scopus in terms of 19,166,705 matched documents. As the result, publication years are identical in 97% of the documents. Numbers of authors are so in 98.8% of the documents. Scopus tends to have a larger number of backward reference IDs. But, MAG includes a larger number of bibliographic data of referenced documents. Spearman's rank correlation coefficient between numbers of forward citations calculated from the two databases (as for documents published in 2005, citations in 10 years window) is significantly high (0.945). On the other hand, affiliation data of all authors are included as for only 43.7 million of documents out of 166 million of MAG documents. MAG is very useful database. However, we might need to use proprietary database depending on research objectives.

1. はじめに

学術論文等の書誌情報に関するデータベースとしては Clarivate 社の Web of Science（WoS）と Elsevier 社の Scopus が計量書誌学の研究などで幅広く活用されてきた。また、Google 社は 2004 年から書誌情報を検索するためのウェブサービスである Google Scholar（GS）を展開しており、研究活動における先行文献調査などで欠かせないツールになっている。これらに加えて、Microsoft 社が 2015 年 6 月に書誌情報や引用情報を検索できるウェブサービス Microsoft Academic Graph[1]（MAG）を公開した（Sinha et al.（2015））。さらに、2018 年 1 月には Digital Science 社も学術文献の書誌情報を検索できる Dimensions のサービスを開始しており、近年は計量書誌学などの分野において活用可能な書誌情報データベースの選択肢が多様になった。

学術文献の書誌情報は、研究活動における先行研究調査のためだけではなく、政府の研究支援のための政策評価や大学ランキングの作成、研究者の評価など様々な用途で利用されている。それらの基盤となる情報であるため、各データベースの特徴を比較した研究が公表されるようになっている（Chadegani et al.（2013）、Thelwall（2018）など）。GS と MAG は、どちらもウェブページをクローリングして収集した情報を活用して構築されたサービスであり、WoS や Scopus と比較しても、どちらも収録文献数が多い[2]。また、MAG はバルクデータが無償で公開されたこともあり、2016 年ごろから文献のカバレッジや情報の正確性を検証した複数の論文が公表されている（表 1）。

例えば、Paszcza（2016）は MAG、WoS、Scopus、GS に収録されているデータ項目や総文献数などを比較して各データベースの特徴をまとめている。Harzing（2016）、Harzing and Alakangas（2017a，2017b）、Hug and Brändle（2017）では、あらかじめ学術文献のリストを用意して、そのリストの文献の MAG、WoS、Scopus、GS の各データベースにおける収録状況を比較することで MAG のカバレッジを検証している。これらの研究では、概ね MAG の文献カバレッジは GS に次ぐ広さであり、WoS を上回ると評価されている。MAG と Scopus のカバレッジの広さの大小関係は、注目したサンプルによって異なる結果になっている。書誌情報の正確性を検証した研究も多数あり、比較対象として用いるデータベースの文献を MAG の文献と DOI で接続して、文献単位で書誌情報の精度を比較する方法をとっている分析が多い（Thelwall（2017；2018a；2018b）など）。書誌情報のなかでは、特に引用情報に注目した研究が多い。例えば、Thelwall らは論文の質を測る指標としての前方引用数（被引用数）をできるだけ早く把握したい場合には、どのデータベースを利用するのが適切か検証するために MAG と Scopus 等との比較分析などを行っている（Thelwall（2017；

[1] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

[2] Microsoft Academic（https://academic.microsoft.com）では、2018 年 9 月時点で約 2 億件の文献が検索可能であるとされている。

2018a；2018b))。Thelwall（2018a）の分析結果によると、前方引用数の平均値は MAG では 6.85、Scopus では 6.74 とかなり近い値であった。また、それぞれのデータベースから作成した前方引用数には高い相関があること（Spearman 順位相関係数 0.948）を示した。Herrmannova and Knoth (2016)や Hug and Brändle (2017)では、論文の出版年や著者数についても精度検証を行っており、MAG の情報の精度が高いことを報告しているが、その一方で MAG における著者の所属情報は欠損が多いことも指摘している（Herrmannova and Knoth（2016））。表 1 に示した先行研究のなかでは Herrmannnova and Knoth (2016)が最も大規模な検証を行っており、MAG、CORE、Mendeley の 3 つのデータベースが収録する 126 万件の文献を DOI で相互に接続して、出版年や引用件数などを比較している。また、Scimago Journal & Country Rank や Webometrics Ranking of World Universities のランキングと MAG で作成したランキングについても比較している。その結果、MAG の引用情報は信頼性のあるソースとして利用できると評価している。

（表 1）

しかしながら、これらの研究もデータベース全体の状況を代表する情報を提供しているとは言えず、より大規模に検証する必要があるだろう。

本論文では Open Academic Society[3]のウェブサイトにおいてバルクデータとして提供されている MAG の全データを用いて Scopus の情報と比較しながらデータの特性を評価する。なお、我々が用いる Scopus のデータは科学技術・学術政策研究所が 2014 年度に購入したバルクデータに基づくものである[4]。

本論文の構成は以下のとおりである。第 2 節では MAG の概略と我々が用いた MAG データについて説明する。第 3 節では MAG と Scopus データの論文単位で接続した方法と比較分析のためのサンプルについて説明し、出版年、著者数、参考文献数、および前方引用数について MAG と Scopus を比較した結果を報告する。第 4 節では、著者情報と著者所属情報のカバレッジと利用可能性を検証した。第 5 節に結論をまとめる。

[3] https://www.openacademic.ai

[4] Scopus データの利用にあたっては、文部科学省科学技術・学術政策研究所の科学技術・学術基盤調査研究室の伊神正貫氏に大変お世話になった。あらためて、ここに感謝の意を表したい。本論文に残された誤りはすべて著者の責任に帰するものである。

2. Microsoft Academic Graph

2.1. Microsoft Academic Graph について

Microsoft Academic Graph の原型は、Microsoft 社が 2006 年に開始したウェブサービスである Windows Live Academic（WLA）にさかのぼる（Ortega（2014））。WLA は ScienceDirect や Wiley Online Library などの複数の出版社のオンライン・ジャーナルや学会論文集などを横断的に検索できるサービスとして開始されたが 2008 年には終了した。しかし、その後、Microsoft Academic Search（MAS）として再開されて、学術文献の書誌情報や引用情報をより大規模に検索することができるようになった。Ortega (2014) によると、MAS は Elsevier や Springer などのオンライン・ジャーナル、論文リポジトリー・サイトの arXiv.org やデジタルオブジェクト識別子(DOI)の公式登録機関のひとつである CrossRef[5,6] などから得られる書誌情報とともに、ウェブのクローリングによって収集された書誌情報も活用してバックデータを構築して検索サービスを提供していた。Microsoft Academic Graph は、MAS を前身として、2015 年 6 月にサービスが公開され、2017 年 7 月に正式サービスが開始された。

Microsoft Academic Graph のデータにアクセスする方法はいくつかある。Microsoft Academic[7] のサイトにおいてキーワードなどで検索をする方法の他に、Academic Knowledge API[8]を使ってアクセスする方法が用意されている。また、パワーユーザーには Azure Data Lake Store を通じた利用が勧められている。Microsoft Academic Graph はクローリングによるデータ収集や収録文献数の多さなどの点において GS と比較されるが、API を利用したデータダウンロードに対応していることは Microsoft Academic Graph の有用な特徴である。しかし、今回、我々が用いるのは Open Academic Society のウェブサイトにおいて提供されているバルクデータである。

Open Academic Society では学術論文等の書誌情報についての 2 種類のバルクデータが公開されている。1 つは Microsoft Academic に基づくデータベースであり、特定の時点において取得されたスナップショットデータである (以降では、このバルクデータベースを指して MAG と呼ぶ)。もう一つは、清華大学の研究者が中心となって作成した書誌情報の検索サービスを基礎としたデータベースの AMiner である（Tang et al. 2008)。この二つのデータベースに収録されている文献はかなり重複しており、二つのデータベースの文献を接続するためのリンクテーブルも併せて公表されている。

---

[5] https://www.crossref.org/

[6] van Eck et al.（2018）は Crossref、WoS、Scopus の引用データの収録状況などを比較している。

[7] https://academic.microsoft.com/

[8] https://labs.cognitive.microsoft.com/en-us/project-academic-knowledge

2.2. MAG データの入手と準備

ダウンロードしたデータは Open Academic Society のウェブサイトに 2017 年 6 月 9 日に公表された ZIP 形式で圧縮された合計 102GB のファイルである。データファイルの文字列符号化形式は UTF- 8 であり、JSON 形式で記述されている。これらを Perl で記述したスクリプトで処理して各文献の書誌情報を取り出した後に、リレーショナル・データベース・マネジメント・システム MySQL、及び統計分析用ソフトウェア Stata に読み込んでデータの加工と集計を行った。

2.3. MAG の特徴

我々が用いる MAG には 166,192,182 件の文献が収録されている。最も古い出版年の文献は 1800 年であるが、1980 年代から収録文献数が大きく増加しており（図 1)、1990 年以降の文献が全体の 78%を占める。出版年別の収録文献数は 2014 年のデータが 829 万件で最も多い（後述の表 4 参照)。データベースにはトランケーションがあり、2016 年から収録件数が大きく減少している（2016 年：740 万件、2017 年：265 万件)。

（図 1）

MAG に収録されているデータ項目は、表 2 に示したとおりである。文献タイトル、出版年、著者名については、ほぼ全てのレコードに情報が存在するが、それ以外のデータ項目については欠損も多い。著者名、著者の所属組織、キーワード、研究分野コード、参考文献、データソースの URL の項目については 1 つの文献に複数のレコードがある。

（表 2）

各文献が掲載されたジャーナルや学会論文集の名称は Venue のカラムに収録されており、Venue の情報は 61,051,941 件の文献、全体の 37%に収録されている。Venue の情報は、アルファベットの大文字と小文字の違いを無視すると、24,974 件の表記パターンがあった。学術論文のジャーナルなどの逐次刊行物を識別する ISSN は全て欠損しており、また、Venue の 24,974 件の表記パターンは表記ゆれを含むため、ジャーナルの名称を利用したい場合は、Venue の情報を整理する必要がある（詳細は第 4 節)。

表 3 に文献タイプの情報をまとめた。文献タイプが識別されているレコードは全体の 35%であった。そのうち Journal が 88%、Conference が 7.5%、Book Chapter が 4%であ

り、MAG に収録されている文献はジャーナル掲載論文が中心である。

（表 3）

デジタルオブジェクト識別子[9]（Digital Object Identifier: DOI）は全体の 41%、68,206,107 件のレコードに付されている。DOI は個々の文献を識別するためのコードである。第 3 節において MAG と Scopus の文献を比較する際には、二つのデータベースの文献を DOI で接続して比較分析のためのサンプルを構築する。

言語の分類（lang）は全体の 85%の文献に情報があり、英語の文献が 8,680 万件で全体の 52%を占める。次いで、日本語の文献が多く、1,212 万件（7.3%）ある。また、スペイン語 576 万件（3.5%）、中国語 563 万件（3.4%）、フランス語 449 万件（2.7%）、ドイツ語 251 万件（1.5%）などの文献も多い。ただし、例えば、lang = ja だったとしても論文タイトルや著者名が日本語以外の言語である場合も散見される。

Field of Study（FOS）は MAG 独自の研究分野の分類データであり、文献の Keyword などを基に作成され、論文単位で付与されている、階層的な構造をもつ分類である。最上位の Level 0 は 19 分類[10]であり、最も細かい Level 3 では 5 万件以上の分類になっている。我々が用いた MAG データには Level 3 のコードまで収録されている。分類は逐次アップデートされているため長期的な時系列比較などには向かないとの指摘もある（Hug et al.（2017））。

3. Scopus との比較

3.1. 比較分析のサンプルについて

本節では、MAG に収録されている文献のデータ特性を評価するために、MAG と Scopus のレコードを DOI で接続し、接続できた文献について、各データベースに収録されている書誌情報（出版年、著者数、参考文献数、被引用数）を比較する。利用する Scopus データは科学技術・学術政策研究所が 2014 年度にバルクで購入したもので、主として 1996 年から 2014 年に発行された合計 34,961,473 件の文献を収録している（表 4 ）。この Scopus データのうちで、DOI の情報がある文献は約 60%の 20,985,615 件である。Scopus は 2009 年

[9] DOI の仕組みは 2000 年に開始された。International DOI Foundation (IDF)が開発、管理している。

[10] Academic Knowledge API で 2018 年 8 月 31 日に Level 0 のリストをダウンロードした結果によると、最上位の分類は、Art, Biology, Business, Chemistry, Computer Science, Economics, Engineering, Environmental Science, Geography, Geology, History, Material Science, Mathematics, Medicine, Philosophy, Physics, Political Science, Psychology, and Sociology の 19 分類である。

以降では 70%以上の文献に DOI がついているが、古い文献ほど DOI 情報がない文献[11]が多く、2000 年の文献では約 3 割にとどまる。一方、MAG については、表 4 に示した 1996 年以降の期間では、DOI が付いている文献の比率は 40%前後で推移している。

DOI はインターネットに公表されている論文などのデジタルオブジェクトを恒久的に一意に識別するためのコードであるため、本来は重複することはない。しかし、MAG にも Scopus にも、複数の文献に同じ DOI が付されているケースがある[12]（例えば、あるジャーナルの文献全てに同じ DOI が付されているケースなど）。これらは比較分析のためのサンプルから除外するものとする。MAG と Scopus のレコードが 1 対 1 で接続できた文献は 19,166,705 件であり、これらを MAG と Scopus の比較分析のためのサンプルとする。

1996 年から 2015 年の期間において、このサンプルは Scopus の DOI 付きの文献の 91%をカバーしており、MAG の DOI 付きの文献の 46%を占めている。なお、DOI で接続した文献が間違っている可能性については未検証であるため、今後さらに精査する必要があると考えられる。

（表 4）

比較分析サンプルの学術分野別の文献数を表 5 に示した。分野分類は Scopus に収録されている All Science Journal Classification[13]（ASJC）の 2 桁コードを用いた。

比較分析サンプルにおいて文献数が多い学術分野は、27 Medicine（20.6%）、22 Engineering（10.5%）、13 Biochemistry, Genetics & Molecular Biology（9.3%）、31 Physics and Astronomy（8.7%）、17 Computer Science（6.4%）などであり、この 5 つの分野でサンプルの 56%を占める。学術分野別で Scopus 全文献に対して比較分析サンプルに含まれた文献の比率が高かったのは 28 Neuroscience（77.2%）や 13 Biochemistry, Genetics & Molecular Biology（71.0%）、18 Decision Science（70.5%）などであり、逆に、比較分析サンプルへの収録率が低かった分野は 12 Arts and Humanity（35.2%）、34 Veterinary（37.8%）、14 Business,

---

[11] ジャーナルのウェブサイトで確認してみると、Scopus には DOI が収録されていないが、実際には DOI が付与されている文献も存在する。DOI の制度は 2000 年に始まったものであり、過去にさかのぼって DOI の付与を行っているジャーナルも存在する。上記の DOI 収録率は 2014 年に購入した Scopus データに基づくものであり、新しい Scopus では DOI 情報の収録率はもっと高いものと思われる。

[12] MAG において、DOI 付きの文献 68,206,107 件のうち、重複した DOI が付された文献は 775,121 件（1.14%）あった。Scopus では、DOI 付きの文献 20,985,615 件のうち、重複した DOI が付された文献は 445,962 件（2.13%）あった。

[13] 本来は、ASJC はジャーナル単位の分類であり文献単位の分類ではない。多くの場合、複数の 4 桁コードが各ジャーナルに付されている。ここでは、文献ごとにランダムに 2 桁コードを 1 つ選択して集計に用いた。

Management & Accounting (40.2%)などである。

(表 5)

3.2. DOI 接続データを用いた比較

3.2.1. 出版年、著者数の比較

文献の出版年、著者名や参考文献は書誌情報を用いた研究において基本情報として重要である。MAG と Scopus の出版年を比較した結果を表 6 に示した。比較分析サンプルにおいて 97%の文献は同じ出版年であった。また、文献ごとの著者数は 98.8%の文献で同じであった（表 7)。MAG の出版年と著者数の情報の精度はかなり高いといえるだろう。

(表 6)

(表 7)

3.2.2. 後方引用数（参考文献数）の比較

計量書誌学において論文の後方引用文献（参考文献）の情報は、論文の前方引用数（被引用数)、雑誌のインパクトファクター、研究者の h-index を作成するときや引用ネットワークの分析など、さまざまに利用される。したがって、前方引用数などの分析を行う前に、まずそれらの基になるデータベースの後方引用文献の情報の特性について検証しておくことは重要である。

MAG と Scopus の後方引用数を比較した結果を先にまとめておくと、主に以下のような傾向を指摘できる。① MAG よりも Scopus の方が後方引用数は多い。② 質の高いジャーナルでは MAG の後方引用文献の収録率は高い。③ それぞれのデータベースの書誌情報とリンクされている後方引用文献だけに注目すると MAG の方が Scopus よりも後方引用数は多い。④ 期間を限定して集計すると、両データベースにおける後方引用数は似た水準になる。

Scopus の場合は基本的には後方引用文献リストの全ての文献に参考文献 ID が付されて収録されている。後方引用文献の情報が存在する場合は、参考文献 ID の数は実際の後方引用数であると考えて問題ないと思われる[14]。ただし、参考文献 ID が付されていても、その

[14] 入手可能なジャーナルについて合計 20 件ほどの文献をランダムに選んで、実際の論文に引用されている後方引用数と Scopus に収録されている参考文献 ID の数を目視で比較してみたが、確認した範囲では後方引用数は正しかった。

文献の書誌情報が Scopus に収録されているとは限らない。つまり、参考文献 ID が存在しても、その文献がどのような論文か分からないレコードがある。一方、MAG の場合は、書誌情報が MAG に収録されている文献のみが後方引用文献として ID が付されて収録されており、このデータ収録方針の違いには注意する必要がある。この点は Herrmannova and Knoth (2016)や Haunschild et al.（2018）などでも指摘されている。

MAG と Scopus の後方引用文献の有無について集計した結果を表 8 に示した。どちらのデータベースにも後方引用文献情報が全く含まれていない文献は全体の 2.4%である。Scopus には後方引用文献情報があるが MAG にはない文献は 242 万件（12.7%）ある。その逆のケース、つまり Scopus には後方引用文献情報がないが MAG にはある文献も 40 万件（2.1%）存在している。どちらのデータベースも後方引用文献情報が必ずしも完備ではない。

（表 8）

MAG の場合はクローリングで収集できていない情報があるためと思われる。ウェブページの構造はジャーナルごとに、またはオンライン・ジャーナル出版社ごとに定形化されており、クローリングで収集できる情報とできない情報があるだろう。図 2 は、ジャーナルの質（SCIMAGO データが提供する Q1〜Q4 の 4 分類[15]）によって、ジャーナルごとに後方引用文献情報の収録率がどのように異なるか傾向をみたものである。ジャーナルの質は Q1（インパクトファクターが最も高いグループ）〜Q4（同じく最も低いグループ）の 4 つのカテゴリーに分かれている。最も質の高い Q1 のジャーナル 6,878 誌のうちで、5,070 誌（74%）については 90%〜100%の文献に後方引用文献情報が存在する。ランクの高いジャーナルの方が後方引用文献情報の収録率がよい傾向にあることが分かる。また、図 3 には、ジャーナルのランクごとに、後方引用文献情報がある文献の平均比率の推移を示した。2000 年代半ば以降に発行された Q1 ジャーナルでは平均して約 90%の文献に後方引用文献情報がある。

（図 2）

（図 3）

次に、MAG と Scopus の後方引用数の違いをみてみる。参考文献数がゼロである場合は

[15] Scopus データベースにはジャーナルの ISSN 情報が収録されている。ここでは、比較サンプルの MAG 文献に Scopus の ISSN を接続して、その ISSN を用いて SCIMAGO の 13000 誌に Q1〜Q4 のランクを接続した。

参考文献の情報が欠損しているとみなした。両方のデータベースに少なくとも 1 件の後方引用文献情報がある文献（全体の 82.8%）に注目すると、平均後方引用数は Scopus が 33.1 で MAG は 27.4 であり Scopus の方が多い（表 8 ）。図 4 には、後方引用数の差の分布と、後方引用数の散布図を示した。この件数の差は、MAG と Scopus の後方引用文献情報の収録方針の違いによる影響が大きいと考えられる。

（図 4）

MAG は、前述のとおり、書誌情報がデータベース内に存在する後方引用文献だけが収録されている。そこで、Scopus でも同様に書誌情報が Scopus に含まれている後方引用文献に限定して比較してみる。

表 9 には、それぞれのデータベースに書誌情報が存在する後方引用文献に限って、後方引用文献情報の有無を集計した結果を示した。1996 年〜2015 年の比較分析サンプルでは、両方のデータベースに少なくとも 1 件の後方引用文献情報がある文献は 78.7%（15,084,033 件）存在し、MAG の平均値は 28.4 件、Scopus では 19.5 件だった。書誌情報が存在する後方引用文献に限ってカウントしているので、表 8 とは異なり、収録期間が長くデータベース全体の規模が大きい MAG の方が後方引用数は多い結果になった。Scopus は主に 1996 年以降に出版された文献しか収録されていないため、書誌情報がある後方引用数はデータベースの左側トランケーションの影響が大きく、特に 1990 年代の Scopus 文献では平均後方引用数が小さい。

（表 9）

特定の年に出版された文献に注目して後方引用文献の出版年をコントロールして集計した値を MAG と Scopus で比較することで、データベースの収録範囲の違いによる影響について考察する。ここでは 2005 年に出版された文献に注目して、書誌情報がある全ての後方引用文献、過去 10 年以内（1996〜2005 年）及び過去 5 年以内（2001〜2005 年）に出版された書誌情報がある後方引用文献だけに限ってカウントした後方引用数を MAG と Scopus それぞれで集計して比較し、また、ピアソンの相関係数とスピアマンの順位相関係数を求めてみる。

書誌情報がある全ての後方引用文献を使って集計した結果を図 5 (a)、過去 10 年以内に出版された後方引用文献に限った結果を図 5 (b)、過去 5 年以内に限った結果を図 5 (c)に示した。図 5 (a)では、MAG の方が後方引用数は大きく、差（=MAG−Scopus）の分布は大きく右に歪んでいる。引用の期間を過去 5 年にコントロールした図 5 (c)では MAG と Scopus の後方引用文献数の差はほとんどなくなる（平均値は 0.35、中央値は 0)、ほぼ左右

対称の分布になった。また、相関係数の値も高まった（ピアソンの相関係数 0.9262、スピアマン順位相関係数 0.8768）。

（図 5）

3.2.3. 前方引用数の比較

論文の質を測る指標として前方引用数（被引用数）がよく用いられる。本節ではそれぞれのデータベースでカウントした前方引用数を比較する。Microsoft Academic Graph の Academic Knowledge API を用いる場合は 2 種類の引用数の情報をダウンロードできる（CC: Citation Count と ECC: Estimated Citation Count）。先行研究において MAG の引用数を検証した論文では CC の情報を使っているケースが多く、CC は MAG に収録されている後方引用文献情報を基にカウントされた値である（Harzing 2016；Harzing and Alakangas 2017a, 2017b；Hug and Brändle 2017 など）。我々は Open Academic Society からダウンロードした MAG バルクデータの後方引用文献情報を基にして独自にカウントした前方引用数を用いて以下の分析を行う。

既に述べたように、Scopus は 2014 年度に購入したバルクデータを、MAG は 2017 年 6 月におけるスナップショットデータを用いており、二つのデータベースは文献収録期間が異なる。データベースの右側トランケーションの違いを考慮して、論文の出版後 3 年以内、5 年以内、7 年以内、10 年以内に引用されたデータを用いてカウントした前方引用数、及び特に出版後経過年数を考慮しない前方引用数を作成した。

図 6 は、2005 年に出版された文献について、MAG と Scopus でそれぞれ作成した前方引用数を散布図にしたものである。2005 年に出版後、3 年以内、5 年以内、7 年以内の前方引用数は概ね対角付近にプロットされており、MAG と Scopus であまり差がないことが分かる。しかし、今回利用した Scopus データは 2014 年から文献収録数が少なくなるため、10 年以内の前方引用数は MAG でカウントした場合の方が前方引用数の値が大きい傾向にある。しかし、それぞれのデータベースから作成した前方引用数の相関はいずれも高い。出版後 10 年以内に引用された前方引用数ではピアソンの相関係数は 0.9625、スピアマンの順位相関係数は 0.9456 であり、書誌情報分析において利用実績の多い Scopus と比較しても MAG の信頼性は高いといえるだろう。

（図 6）

4. 論文掲載誌と著者情報のカバレッジ

論文データベースを有効に活用するためには、各論文の掲載誌に関する情報（論文の学術分野やインパクトファクターから見た質に関する情報）や論文著者に関する情報を整理することが必要である。ここでは、この両者について MAG の利用可能性について評価した。

4.1. 論文掲載誌情報の評価

論文掲載誌については、MAG のオリジナルデータにおいてジャーナル名（Venue）の記載情報があるが、これに ISSN を付与することで論文の学術分類や学会誌の学術ランキング情報と接続することが可能となる。従って、ここではジャーナル名のテキスト情報を ISSN に変換する作業を試みた。

具体的には、Scopus 掲載誌リスト（Elsevier 社ウェブサイトからダウンロード、2018 年 4 月現在のリスト）における学術誌名と ISSN 対応表を用いて、当該データの学術誌名と MAG データから得られた学術誌名を接続することを試みた。なお、Scopus 掲載誌リストには Scopus の ID 数として 37,062、ISSN 数としては 47,618 の学術誌が収録されている（同じ学術誌についても紙媒体と電子媒体で異なる ISSN が付与されるため 1 つの学術誌に対して複数の ISSN が存在しうる）。接続方法は単語（Token）レベルの Approximate Matching を行った。具体的には単語の頻出頻度の対数値の逆数をウェイトとした Jaccard 指数で 0.8 以上のものを同じ雑誌であるとみなした。

論文総数である 166,192,182 のうち、オリジナル情報において何らかのジャーナル名情報（Venue 情報）が存在するものが 61,051,921（それ以外は当該情報が Null）であり、そのうち 51,401,398 については ISSN 情報が得られた。また、この内容を MAG データと Scopus データを論文の DOI 情報で接続したデータ（第 3 章参照、19,166,705 本）を用いて、Scopus における ISSN 情報をどの程度カバーしているか調べた。その結果、15,355,987 本（全体の約 8 割）については MAG から ISSN 情報が得られることが分かった。このように MAG 全体から見ると、ISSN 情報を付与できた論文数が 1 ／ 3 以下となるが、Scopus 収録論文について見るとかなりの割合の論文について、MAG の情報によって代替することが可能であることが分かった。

4.2. 論文著者の所属機関情報

論文著者の所属機関の情報は、論文数の国別、機関名別推移といった学術情報を用いた基礎的な統計データ処理を行う上で重要である。MAG においては、著者の氏名情報と所属機関の情報は別のレコードとして与えられている。しかし、所属機関については、機関名と機関の所在地情報が混在するテキスト情報となっており、ここから分析上有益な情報を取り

出すことが必要である。

所属機関に関するテキストから所在地や所属機関名に関する情報を取り出す方法については、Stanford Named Entity Recognition System（Stanford NER）を用いた。Stanford NER は Stanford 大学の自然言語処理グループが提供する Named Entity Recognition（NER）システムである(https://nlp.stanford.edu/ner/)。なお、NER システムは Stanford 大学のものの他、spaCy, LingPipe, Python-NLTK などの各種ツールが存在するが、それらの中で Stanford NER は比較的良好なパフォーマンスを示すことが分かっている（Jiang et al. 2016)。ここでは所在地情報（Stanford NER は Country, Province, City の 3 種類の地理情報を抽出する）を用いて機関の所在地である国コードを作成した。

表 10 は、著者の氏名情報と所属機関の情報の有無について見たものである。（論文毎に）すべての著者について情報が存在（Yes)、一部の著者について存在（Partly Yes）及びすべての著者について存在しない（No）の 3 通りでそれぞれの論文数を示している。

まず、氏名の情報については約 1.66 億本のほとんどの論文において存在する（少しでも著者氏名が欠けている論文数は 2,000 件程度である)。一方で、著者の所属機関情報については多くの論文において情報が存在しないことが分かった。すべての著者において所属機関情報が存在する論文が約 4,374 万件、一部の著者について所属機関情報が存在する論文が約 290 万件で残りの 1.2 億万件については機関情報なしとなっている。なお、一部の著者情報が欠けている論文数割合は非常に小さいので、所属機関情報の有無は論文ごとにほぼ決まっている。

（表 10）

ただし、SCIMAGO の対象論文のみをみると所属機関情報がある論文割合は総数約 5,483 万件のうち、約 2,781 万件と半数以上になる（表 11)。更に、Stanford NER の機関所在地情報から判別した国コードの付与情報については、何らかの機関情報が存在する約 2,781 万件のうち、約 2,100 万件つまり 3 ／ 4 の機関情報から国コードが判別できたこととなる。しかし、総論文件数 5,483 万件と比較すると半数以下となり、所属機関に関する情報がそもそも欠損値となっていることが大きな制約となっている。

（表 11）

MAG はウェブページ上の情報を定期的にクローリングすることで作成されているので、所属機関の情報が欠損値となっているのは、ウェブページ上の表示形式に問題があることが原因であると考えられる。ウェブページ上の表示形式は学術誌によって統一されているはずなので、情報の欠損状況は学術誌ごとに決まってくる可能性が高い。図 7 は論文の質

（SCIMAGO データが提供する Q1〜Q4 の 4 分類）によって、所属機関の欠損状況をみたものである。Q1（インパクトファクターが最も高いグループ）〜Q4（同じく最も低いグループ）のそれぞれについて、学術誌ごとに国コードの付与割合分布（10 分位）を見たものである。例えば、Q1 から Q3 の学術誌については、トップ 10%の平均付与割合は 95%程度となっており、逆に Q4 については約 4 割の学術誌について付与割合が 0 となっている（P60 の段階で 0%）。インパクトファクターの高い論文において、所属機関情報の利用可能性が高く、質の高い論文に限定することで MAG の利用可能性が高まることを示唆している。

（図 5）

5. まとめ

本論文においては、Microsoft 社が収集した書誌情報データベースの Microsoft Academic Graph（MAG）の利用可能性について、Elsevier 社の商用データベース Scopus をベンチマークとして評価した。MAG の評価について、一部の機関に所属する研究者の著作物について、Scopus の他、Web of Science（WoS）、Google Scholar などの他の書誌情報データベースと比較する論文は公表されているが、データベース全体を対象とした分析は行われていなかった。そこで、今回はすべての論文（約 1.66 億本）を対象にデータベース全体とした定量的分析を行った。

MAG と Scopus のそれぞれに収録されている論文について、DOI でマッチできる同一論文について論文出版年及び著者数について比較したところ、前者については 97.0%、後者については 98.8%の論文において一致した。また、後方引用数については全体的に Scopus の方が大きくなるが（論文 1 本あたりの平均引用数は MAG が 27.4、Scopus が 33.1）、データベースに収録されている書誌の引用に限ると MAG の方が大きくなる（MAG が 28.4 、Scopus が 19.5）。これは MAG が Scopus では収録されていない 1990 年代以前の論文もカバーしているからである。つまり、後方引用論文に関する分析を行う上では、当該論文の書誌情報をより多く有している MAG の方が利用価値が高いということになる。また、前方引用（被引用）について、MAG と Scopus のそれぞれでみた前方引用数のスピアマン順位相関係数は 0.90~0.95（2005 年出版年の論文）となり、ほぼ同様の精度であることが分かった。

次に論文数で見た研究パフォーマンスの個人別・機関別評価を行うために必要となる論文著者、著者所属機関情報について見た。Scopus は論文出版元からこれらの情報を得ているので、ほぼすべての論文について、上記の情報を得ることができる。一方で、MAG はインターネット上の情報をクローリングして得られたものなので、サイトの構造から上記の

情報を得られない、つまり情報が欠落しているものが多くみられる。著者情報については、ほとんどの論文について得られるものの、所属機関情報は多くの論文において欠落していることが分かった。具体的には、約 1.66 億本のうち、すべての著者において所属機関情報が存在する論文が約 4,374 万件、一部の著者について所属機関情報が存在する論文が約 290 万件で残りの 1.2 億件については機関情報が空欄となっている。これを SCIMAGO 収録論文約 5,483 万件に限ってみると機関情報ありの論文の割合が半数近くまで上昇し、かつその中でも相対的に質の高いジャーナル論文に限るとさらにその割合は上昇する。しかし、その場合でも多くの論文で情報が欠落しており、大学ランキングなどの機関ごとの研究パフォーマンスを評価するための材料として不十分であるといえる。

MAG はバルクデータとして無償で提供されているので、同データが Scopus や WoS などの商用データベースの代替データとして利用できることの意義は大きい。出版年、著者情報及び引用情報については、Scopus とそん色ないレベル（後方引用についてはむしろ MAG の方が有用なケースもあるレベル）のデータであることが分かった。一方で、論文著者の機関名情報については欠落している論文が多く、当該情報を利用する際には注意が必要である。結論として、MAG は全体としては有用なデータベースであるといえるが、所属機関情報を用いる分析など、研究目的によっては商用データベースに頼らざるを得ないというのが現状といえる。

今回は、Scopus との比較をベースに MAG の評価を行ったが、今後の作業として、まず WoS との比較を挙げることができる。Scopus においても書誌情報に誤りがある可能性があり、WoS を加えることでより真の値に近い情報と比較することが可能である。また、MAG の特性についてさらに検証するためには、ジャーナル毎の分析を進めることも有益である。MAG がウェブ情報をクローリングして得られたものであるため、例えば機関情報の欠落などの問題は、ジャーナルや出版社のウェブページの構造に影響されると予想できる。これらの分析を通じてデータベースの特性がより詳細に明らかになることは、今後の計量書誌情報学の発展にとって重要であると考える。

参考文献

Chadegani, A. A., H. Salehi, M. M. Yunus, H. Farhadi, M. Fooladi, M. Farhadi and N. A. Ebrahim (2013) "A Comparison between Two Main Academic Literature Collections: Web of Science and Scopus Databases," *Asian Social Science*, Vol.9, No.5, pp.18-26, DOI: 10.5539/ass.v9n5p18.

Harzing, A. (2016) "Microsoft Academic (Search): a Phoenix arisen from the ashes?" *Scientometrics*, Vol.108, Issue 3, pp.1637-1647, DOI: 10.1007/s11192-016-2026-y.

Harzing, A. and S. Alakangas (2017a) "Microsoft Academic: is the phoenix getting wings?" *Scientometrics*, Vol.110, Issue 1, pp.371-383, DOI: 10.1007/s11192-016-2185-x.

Harzing, A. and S. Alakangas (2017b) "Microsoft Academic is one year old: the phoenix is ready to leave the nest," *Scientometrics*, Vol.112, Issue 3, pp.1887-1894, DOI:10.1007/s11192-017-2454-3.

Haunschild, R., S. E. Hug, M. P. Brändle and L. Bornmann (2018) "The number of linked references of publications in Microsoft Academic in comparison with the Web of Science, *Scientometrics*, Vol.114, Issue 1, pp.367-370, DOI: 10.1007/s11192-017-2567-8.

Herrmannova, D. and P. Knoth (2016) "An analysis of the Microsoft Academic Graph," *D-Lib Magazine*, Vol.22, Number 9/10, DOI: 10.1045/september2016-herrmannova.

Hug, S. E. and M. P. Brändle (2017) "The coverage of Microsoft Academic: Analyzing the publication output of a university," *Scientometrics*, Vol.113, Issue 3, pp.1551-1571, DOI: 10.1007/s11192-017-2535-3

Hug, S. E., M. Ochsner and M. P. Brändle (2017) "Citation Analysis with Microsoft Academic," *Scientometrics*, Vol.111, Issue 1, pp.371-378, DOI: 10.1007/s11192-017-2247-8.

Jiang, R., R. E. Banchs and H. Li (2016) "Evaluating and Combining Named Entity Recognition System," *Proceedings of the Sixth Named Entity Workshop*, join with 54th ACL, 21-27, Berlin Germany, August 12, 2016, DOI: 10.18653/v1/W16-2703.

Ortega, J. L. (2014) Academic Search Engines: A Quantitative Outlook, Chandos Information Professional Series 1st Edition, Chandos Publishing, Elsevier, ISBN 978-1-84334-791-0 (Print), ISBN 978-1-78063-472-2 (Online).

Paszcza, B. (2016) "Comparison of Microsoft Academic Graph with Other Scholarly Citation Databases," Thesis for the Degree of Master of Science, University of Southampton, September 2016, DOI: 10.13140/RG.2.2.21858.94405.

Sinha, A., Z. Shen, Y. Song, H. Ma, D. Eide, B. Hsu and K. Wang (2015) "An Overview of Microsoft Academic Service (MAS) and Applications," *Proceedings of the 24th*

International Conference on World Wide Web (WWW '15 Companion), ACM, New York, NY, USA, pp.243-246, DOI: 10.1145/2740908.2742839.

Tang, J., J. Zhang, L. Yao, J. Li, L. Zhang and Z. Su (2008) "ArnetMiner: Extraction and Mining of Academic Social Networks," *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008)*, pp.990-998.

Thelwall, M. (2017) "Microsoft Academic: A multidisciplinary comparison of citation counts with Scopus and Mendeley for 29 journals," *Journal of Informetrics*, Vol.11, Issue 4, pp.1201-1212, DOI: 10.1016/j.joi.2017.10.006.

Thelwall, M. (2018a) "Microsoft Academic automatic document searches: Accuracy for journal articles and suitability for citation analysis," *Journal of Informetrics*, Vol.12, Issue 1, pp.1-9, DOI: 10.1016/j.joi.2017.11.001.

Thelwall, M. (2018b) "Does Microsoft Academic find early citations?" *Scientometrics*, Vol.114, Issue 1, pp.325-334, DOI: 10.1007/s11192-017-2558-9.

Thelwall, M. (2018c) "Dimensions: A competitor to Scopus and the Web of Science?" *Journal of Informetrics*, Vol.12, Issue 2, pp.430-435, DOI: 10.1016/j.joi.2018.03.006.

van Eck, N. J., L. Waltman, V. Larivière and C. Sugimoto (2018) "Crossref as a new source of citation data: A comparison with Web of Science and Scopus," A blog post in the website of the Centre for Science and Technology Studies (CWTS), Leiden University, URL: https://www.cwts.nl/blog?article=n-r2s234 (Last access: 25 September 2018).

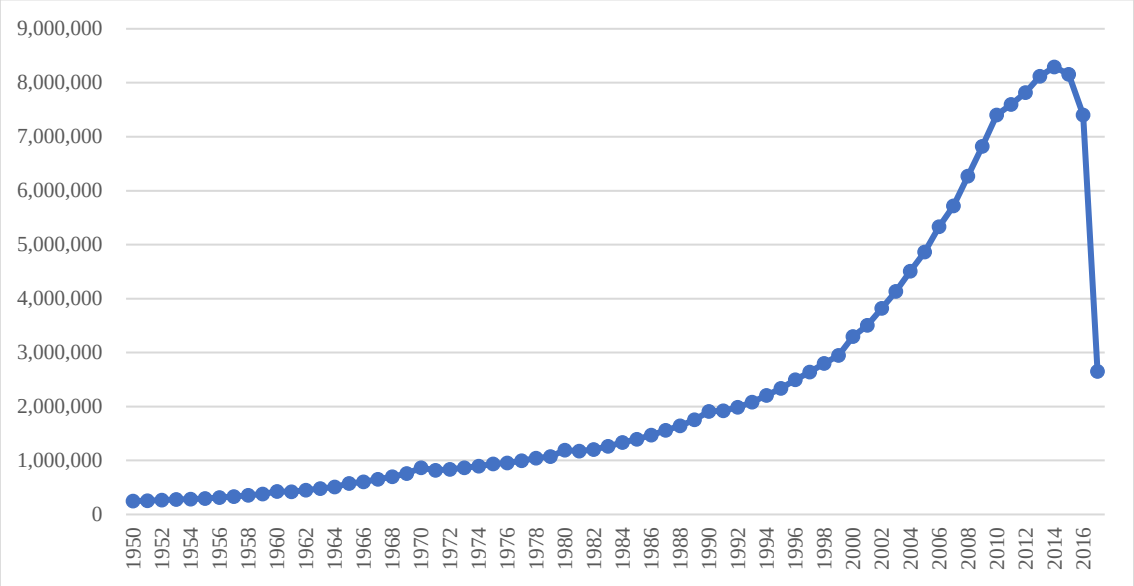
図表

表 1. 先行研究

	データベース	サンプル	主な比較項目
Harzing (2016)	MAG, WoS, Scopus, GS	The author's articles written in 1995-2016.	Coverage, Citation Count.
Herrmannova and Knoth (2016)	MAG, CORE, Mendeley	Intersection set of MAG, CORE and Mendeley 1.26 million documents.	Publication year, Citation Count, Ranking.
Harzing and Alakangas (2017a)	MAG, WoS, Scopus, GS	Articles of 145 academics at the University of Melbourne.	Coverage, Citation Count, Citation indexes
Harzing and Alakangas (2017b)	MAG, WoS, Scopus, GS	The author's articles written in 1995-2016, and articles of 145 academics at the University of Melbourne.	Coverage, Citation Count, Citation indexes
Hug, Ochsner and Brändle (2017)	MAG, Scopus, GS	Three researchers' publications (n = 57).	Citation indexes.
Hug and Brändle (2017)	MAG, WoS, Scopus	Publications included in the Zurich Open Archive and Repository (ZORA): 91,215 documents.	Coverage, Citation Count. Publication year, Number of authors.
Thelwall (2017)	MAG, Scopus, Mendeley	172,752 articles in 29 journals	Citation Count
Thelwall (2018a)	MAG, Scopus	126,312 articles in Scopus 323 subfields in 2012.	Citation Count.
Thelwall (2018b)	MAG, Scopus, Mendeley	44,398 articles in Nature, Science, and 7 journals in library & information science	Citation Count.

(出典：著者)

図 1. 出版年別の収録件数の推移



（出典：著者）

表 2. データ項目

データ項目	説明	Nullでないレコード数	%	関係
id	MAG 文献ID	166,192,182	100%	1 : 1
year	出版年	166,192,182	100%	1 : 1
title	文献タイトル	166,192,182	100%	1 : 1
abstract	要旨	5,593,007	3.4%	1 : 1
publisher	発行者	100,358,932	60.4%	1 : 1
venue	ジャーナル名等	61,051,941	36.7%	1 : 1
doc_type	文献タイプ	58,834,175	35.4%	1 : 1
doi	デジタルオブジェクト識別子	68,206,107	41.0%	1 : 1
lang	言語	141,682,192	85.3%	1 : 1
issn	ISSN	0	0%	1 : 1
isbn	ISBN	0	0%	1 : 1
volume	巻	85,435,560	51.4%	1 : 1
issue	号	83,184,991	50.1%	1 : 1
page_stat	文献開始ページ	98,093,266	59.0%	1 : 1
page_end	文献最終ページ	85,031,970	51.2%	1 : 1
n_citation	引用数	52,833,805	31.8%	1 : 1
authors.name	著者名	166,192,008	99.9%	1 : 多
authors.org	著者の所属組織	46,649,243	28.1%	1 : 多
references	参考文献	47,720,081	28.7%	1 : 多
keywords	キーワード	94,476,176	56.8%	1 : 多
fos	研究分野 Field of Study	109,993,272	66.2%	1 : 多
url	データソースのURL	161,847,144	97.4%	1 : 多

（出典：著者）

表 3. 文献タイプ

doc_type	N	%	（Null以外）%
Book	486,218	0.3%	0.8%
BookChapter	2,330,482	1.4%	4.0%
BookReferenceEntry	103,575	0.1%	0.2%
Conference	4,403,689	2.6%	7.5%
Journal	51,510,211	31.0%	87.6%
(Null)	107,358,007	64.6%	-
Total	166,192,182	100%	100%

（出典：著者）

表 4. MAG と Scopus の文献数

出版年	MAG [A] レコード数	MAG [B] DOIあり	MAG [B/A]	Scopus [C] レコード数	Scopus [D] DOIあり	Scopus [D/C]	1対1で接続できた文献 [E]	1対1で接続できた文献 [E/B]	1対1で接続できた文献 [E/D]
1996	2,499,158	1,072,810	42.9%	1,143,317	259,990	22.7%	239,383	22.3%	92.1%
1997	2,639,614	1,110,842	42.1%	1,170,368	250,357	21.4%	232,834	21.0%	93.0%
1998	2,801,727	1,151,428	41.1%	1,172,220	298,090	25.4%	276,763	24.0%	92.8%
1999	2,946,394	1,172,988	39.8%	1,179,704	358,823	30.4%	332,862	28.4%	92.8%
2000	3,297,105	1,266,712	38.4%	1,243,774	375,600	30.2%	345,538	27.3%	92.0%
2001	3,508,116	1,318,999	37.6%	1,343,833	559,156	41.6%	516,375	39.1%	92.3%
2002	3,818,082	1,389,223	36.4%	1,398,058	638,712	45.7%	590,363	42.5%	92.4%
2003	4,132,570	1,490,084	36.1%	1,473,203	716,320	48.6%	651,436	43.7%	90.9%
2004	4,510,265	1,649,469	36.6%	1,614,021	824,365	51.1%	760,529	46.1%	92.3%
2005	4,861,714	1,755,976	36.1%	1,844,749	1,040,353	56.4%	965,695	55.0%	92.8%
2006	5,335,021	1,953,924	36.6%	1,946,119	1,214,211	62.4%	1,107,738	56.7%	91.2%
2007	5,720,582	2,111,122	36.9%	2,057,504	1,336,698	65.0%	1,231,820	58.3%	92.2%
2008	6,270,317	2,320,546	37.0%	2,157,617	1,473,686	68.3%	1,360,662	58.6%	92.3%
2009	6,821,538	2,536,574	37.2%	2,262,452	1,599,315	70.7%	1,470,783	58.0%	92.0%
2010	7,405,212	2,851,613	38.5%	2,395,921	1,722,923	71.9%	1,585,870	55.6%	92.0%
2011	7,599,908	2,854,822	37.6%	2,544,833	1,873,362	73.6%	1,725,116	60.4%	92.1%
2012	7,816,512	3,065,135	39.2%	2,630,735	1,998,915	76.0%	1,823,749	59.5%	91.2%
2013	8,122,294	3,324,308	40.9%	2,689,588	2,131,696	79.3%	1,925,549	57.9%	90.3%
2014	8,294,382	3,535,456	42.6%	2,454,440	2,087,864	85.1%	1,826,197	51.7%	87.5%
2015	8,158,176	3,556,525	43.6%	233,292	220,187	94.4%	194,701	5.5%	88.4%
小計	106,558,687	41,488,556	38.9%	34,955,748	20,980,623	60.0%	19,163,963	46.2%	91.3%
上記以外の年	59,633,495	26,717,551	44.8%	5,725	4,992	87.2%	2,742	0.0%	54.9%
合計	166,192,182	68,206,107	41.0%	34,961,473	20,985,615	60.0%	19,166,705	28.1%	91.3%

（出典：著者）

表 5. 比較分析サンプルの学術分野について

ASJC	[A] Scopus 全レコード	%	[B] 比較分析サンプル	%	[B/A]
10 General	309,678	0.9%	146,278	0.8%	47.2%
11 Agricultural & Biological Sciences	1,644,680	4.7%	887,307	4.6%	54.0%
12 Arts and Humanities	661,403	1.9%	232,647	1.2%	35.2%
13 Biochemistry, Genetics & Molecular Biology	2,515,809	7.2%	1,786,310	9.3%	71.0%
14 Business, Management & Accounting	479,143	1.4%	192,795	1.0%	40.2%
15 Chemical Engineering	749,061	2.1%	314,775	1.6%	42.0%
16 Chemistry	1,575,037	4.5%	964,698	5.0%	61.2%
17 Computer Science	1,924,791	5.5%	1,227,123	6.4%	63.8%
18 Decision Sciences	114,993	0.3%	81,065	0.4%	70.5%
19 Earth and Planetary Sciences	1,091,266	3.1%	559,296	2.9%	51.3%
20 Economics, Econometrics & Finance	301,178	0.9%	161,328	0.8%	53.6%
21 Energy	493,624	1.4%	214,787	1.1%	43.5%
22 Engineering	4,236,955	12.1%	2,015,196	10.5%	47.6%
23 Environmental Science	875,579	2.5%	507,845	2.6%	58.0%
24 Immunology & Microbiology	614,989	1.8%	419,428	2.2%	68.2%
25 Materials Science	1,612,422	4.6%	879,188	4.6%	54.5%
26 Mathematics	1,049,336	3.0%	680,878	3.6%	64.9%
27 Medicine	7,573,133	21.7%	3,956,127	20.6%	52.2%
28 Neuroscience	553,115	1.6%	427,087	2.2%	77.2%
29 Nursing	325,279	0.9%	162,737	0.8%	50.0%
30 Pharmacology, Toxicology & Pharmaceutics	752,415	2.2%	406,639	2.1%	54.0%
31 Physics and Astronomy	2,533,414	7.2%	1,672,220	8.7%	66.0%
32 Psychology	450,478	1.3%	280,507	1.5%	62.3%
33 Social Sciences	1,492,988	4.3%	703,121	3.7%	47.1%
34 Veterinary	239,408	0.7%	90,508	0.5%	37.8%
35 Dentistry	161,397	0.5%	86,539	0.5%	53.6%
36 Health Professions	207,101	0.6%	108,034	0.6%	52.2%
N/A	422,801	1.2%	2,242	0.0%	0.5%
Total	34,961,473	100%	19,166,705	100%	54.8%

（出典：著者）

表 6. MAG と Scopus の比較：同じ出版年の文献

Scopus 出版年	N	著者数が一致	パーセント
1995以前	266	266	100%
1996	239,383	235,297	98.3%
1997	232,834	228,437	98.1%
1998	276,763	271,135	98.0%
1999	332,862	326,583	98.1%
2000	345,538	339,622	98.3%
2001	516,375	507,027	98.2%
2002	590,363	581,360	98.5%
2003	651,436	641,805	98.5%
2004	760,529	750,004	98.6%
2005	965,695	951,863	98.6%
2006	1,107,738	1,095,688	98.9%
2007	1,231,820	1,217,876	98.9%
2008	1,360,662	1,347,015	99.0%
2009	1,470,783	1,458,172	99.1%
2010	1,585,870	1,572,132	99.1%
2011	1,725,116	1,707,812	99.0%
2012	1,823,749	1,805,245	99.0%
2013	1,925,549	1,905,463	99.0%
2014	1,826,197	1,804,706	98.8%
2015	194,701	192,526	98.9%
N/A	2,476	2,436	98.4%
Total	19,166,705	18,942,470	98.8%

（出典：著者）

表 7. MAG と Scopus の比較：同じ著者数の文献

Scopus Publication year	N	Scopus & MAG: Same No. authors	Percent
1995以前	266	266	100%
1996	239,383	235,297	98.3%
1997	232,834	228,437	98.1%
1998	276,763	271,135	98.0%
1999	332,862	326,583	98.1%
2000	345,538	339,622	98.3%
2001	516,375	507,027	98.2%
2002	590,363	581,360	98.5%
2003	651,436	641,805	98.5%
2004	760,529	750,004	98.6%
2005	965,695	951,863	98.6%
2006	1,107,738	1,095,688	98.9%
2007	1,231,820	1,217,876	98.9%
2008	1,360,662	1,347,015	99.0%
2009	1,470,783	1,458,172	99.1%
2010	1,585,870	1,572,132	99.1%
2011	1,725,116	1,707,812	99.0%
2012	1,823,749	1,805,245	99.0%
2013	1,925,549	1,905,463	99.0%
2014	1,826,197	1,804,706	98.8%
2015	194,701	192,526	98.9%
N/A	2,476	2,436	98.4%
Total	19,166,705	18,942,470	98.8%

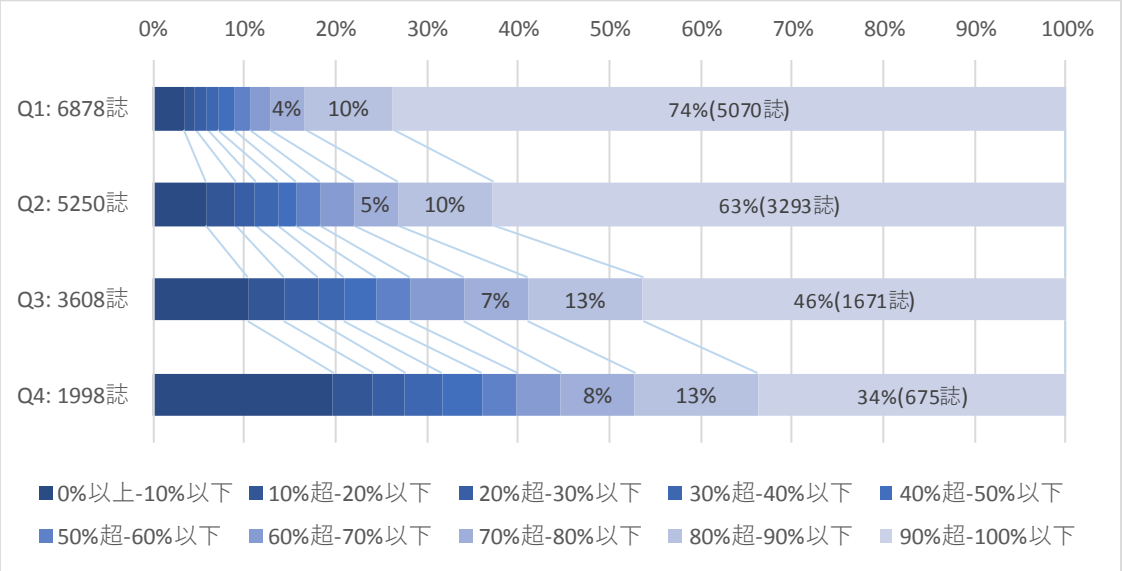
（出典：著者）

表 8. MAG と Scopus：後方引用文献情報の有無、平均後方引用数

Scopus 出版年	N	(A) MAGなし Scopusなし	(B) MAGなし Scopusあり	(C) MAGあり Scopusなし	(D) MAGあり Scopusあり	(D) について Scopus 平均値	(D) について MAG 平均値
1996	239,383	3.3%	19.9%	6.6%	70.2%	27.9	21.1
1997	232,834	3.8%	17.6%	6.7%	71.9%	31.0	24.3
1998	276,763	3.6%	16.0%	6.5%	74.0%	30.9	24.6
1999	332,862	2.7%	13.6%	4.9%	78.8%	30.6	24.1
2000	345,538	2.9%	14.3%	2.8%	79.9%	30.6	24.2
2001	516,375	2.0%	16.8%	0.9%	80.3%	30.2	23.2
2002	590,363	2.4%	16.2%	1.0%	80.5%	31.0	24.0
2003	651,436	2.3%	14.2%	0.8%	82.7%	31.6	25.2
2004	760,529	2.3%	14.6%	0.8%	82.4%	32.0	25.7
2005	965,695	2.6%	12.4%	0.7%	84.2%	31.4	25.6
2006	1,107,738	2.7%	11.8%	0.8%	84.7%	31.5	25.8
2007	1,231,820	2.7%	11.3%	0.8%	85.2%	31.2	25.8
2008	1,360,662	2.8%	10.9%	1.6%	84.7%	31.4	26.3
2009	1,470,783	2.4%	11.4%	0.6%	85.6%	31.8	26.7
2010	1,585,870	2.2%	12.1%	0.5%	85.1%	32.6	27.4
2011	1,725,116	2.1%	12.3%	0.4%	85.2%	33.5	28.2
2012	1,823,749	2.2%	12.9%	0.7%	84.2%	35.1	29.6
2013	1,925,549	2.1%	12.9%	1.7%	83.4%	36.2	30.6
2014	1,826,197	2.2%	11.5%	8.9%	77.4%	37.5	31.9
2015	194,701	2.1%	5.5%	14.8%	77.6%	40.2	34.4
Total	19,163,963	2.4%	12.7%	2.1%	82.8%	33.1	27.4

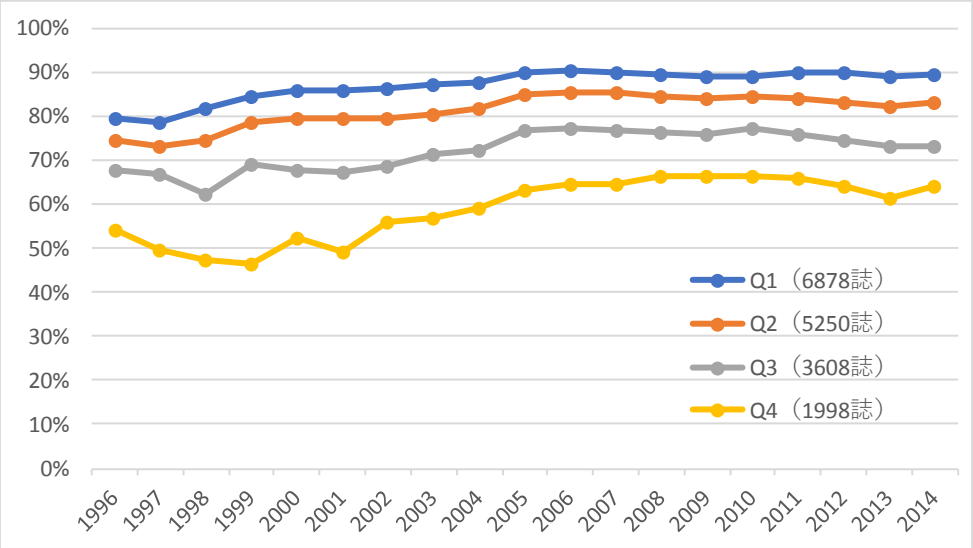
（出典：著者）

図 2. MAG：ジャーナルの質と後方引用文献情報がある文献の比率



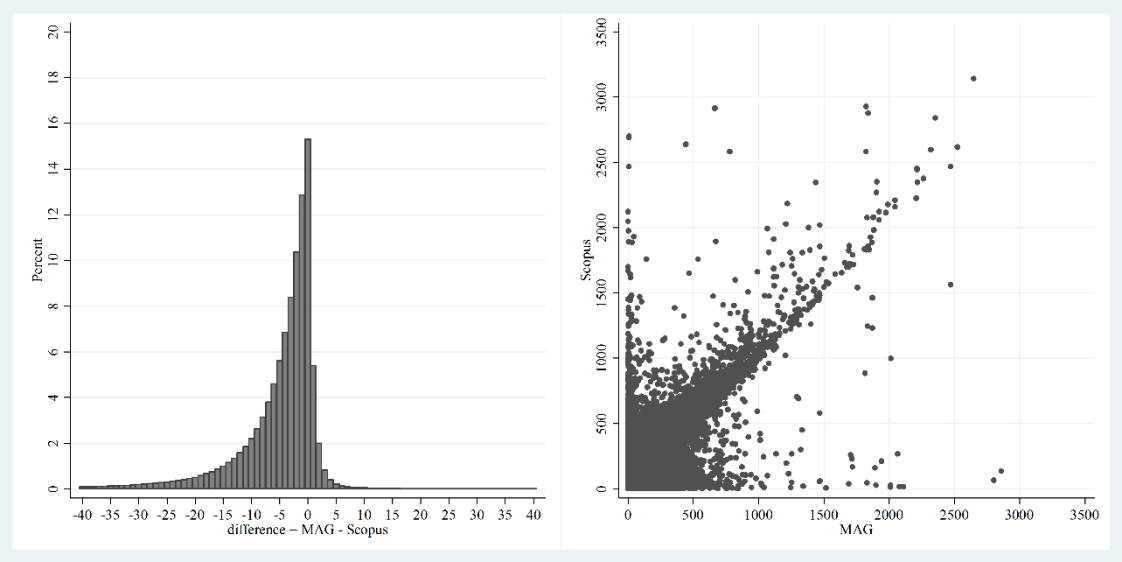
（出典：著者）

図 3. MAG: ジャーナルの質と後方引用文献情報がある文献の平均比率の推移



（出典：著者）

図 4. 後方引用数の差の分布・後方引用数の散布図



（出典：著者）

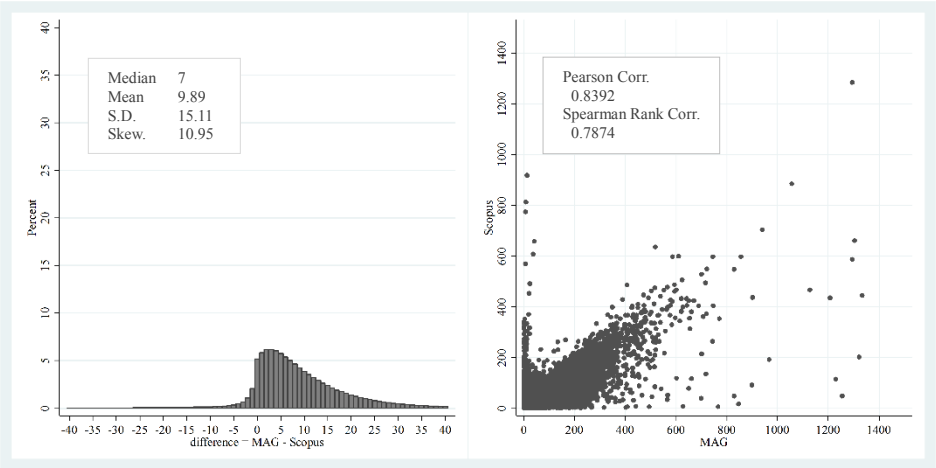
表 9. MAG と Scopus：書誌情報が存在する後方引用文献情報の有無、平均後方引用数

Scopus 出版年	N	(A) MAGなし Scopusなし	(B) MAGなし Scopusあり	(C) MAGあり Scopusなし	(D) MAGあり Scopusあり	(D) について Scopus 平均値	(D) について MAG 平均値
1996	239,383	18.5%	4.7%	65.5%	11.3%	2.0	30.1
1997	232,834	11.3%	10.1%	37.5%	41.2%	3.8	29.2
1998	276,763	7.8%	11.7%	22.7%	57.7%	5.9	27.5
1999	332,862	5.0%	11.3%	16.2%	67.5%	8.0	26.2
2000	345,538	4.9%	12.4%	10.7%	72.1%	9.8	25.7
2001	516,375	4.6%	14.2%	6.5%	74.8%	11.0	24.3
2002	590,363	4.8%	13.7%	5.3%	76.1%	12.8	25.0
2003	651,436	4.4%	12.1%	4.4%	79.1%	14.5	26.0
2004	760,529	4.1%	12.7%	3.6%	79.5%	16.1	26.4
2005	965,695	4.5%	10.6%	4.0%	81.0%	16.5	26.4
2006	1,107,738	4.5%	9.9%	3.9%	81.6%	17.3	26.6
2007	1,231,820	4.8%	9.2%	4.1%	81.9%	17.7	26.6
2008	1,360,662	4.7%	9.0%	4.7%	81.7%	18.5	27.1
2009	1,470,783	4.3%	9.5%	3.5%	82.7%	19.4	27.5
2010	1,585,870	4.3%	10.0%	3.3%	82.4%	20.5	28.1
2011	1,725,116	4.5%	9.9%	3.0%	82.7%	21.8	28.9
2012	1,823,749	4.5%	10.6%	2.6%	82.3%	23.3	30.1
2013	1,925,549	4.4%	10.5%	3.2%	81.8%	24.5	31.1
2014	1,826,197	4.3%	9.4%	10.1%	76.2%	25.8	32.3
2015	194,701	2.7%	4.9%	15.5%	76.9%	28.0	34.7
Total	19,163,963	4.8%	10.3%	6.2%	78.7%	19.5	28.4

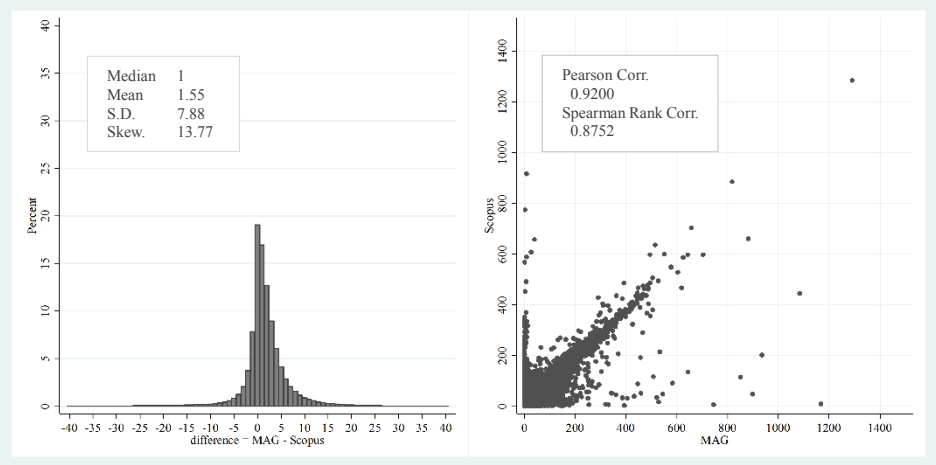
（出典：著者）

図 5. 2005 年出版の文献：後方引用数の差の分布・後方引用数の散布図

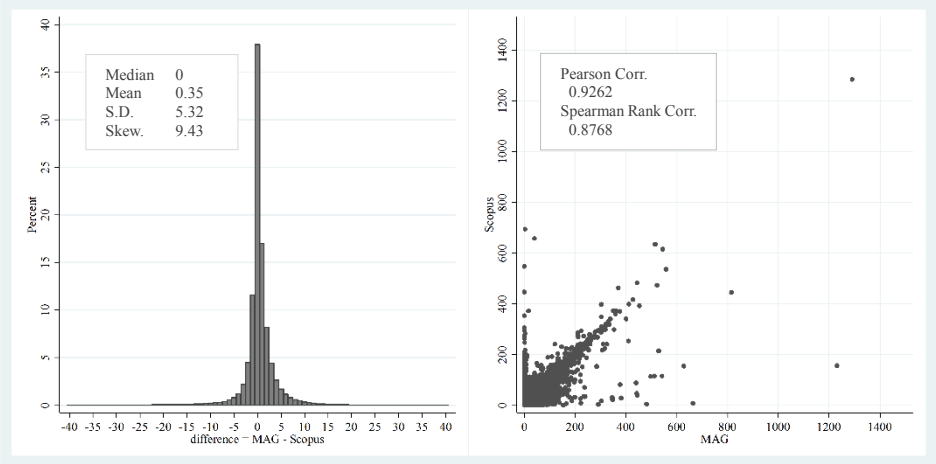
(a) 書誌情報がある後方引用文献：全て



(b) 書誌情報がある後方引用文献：10 年以内の文献のみ

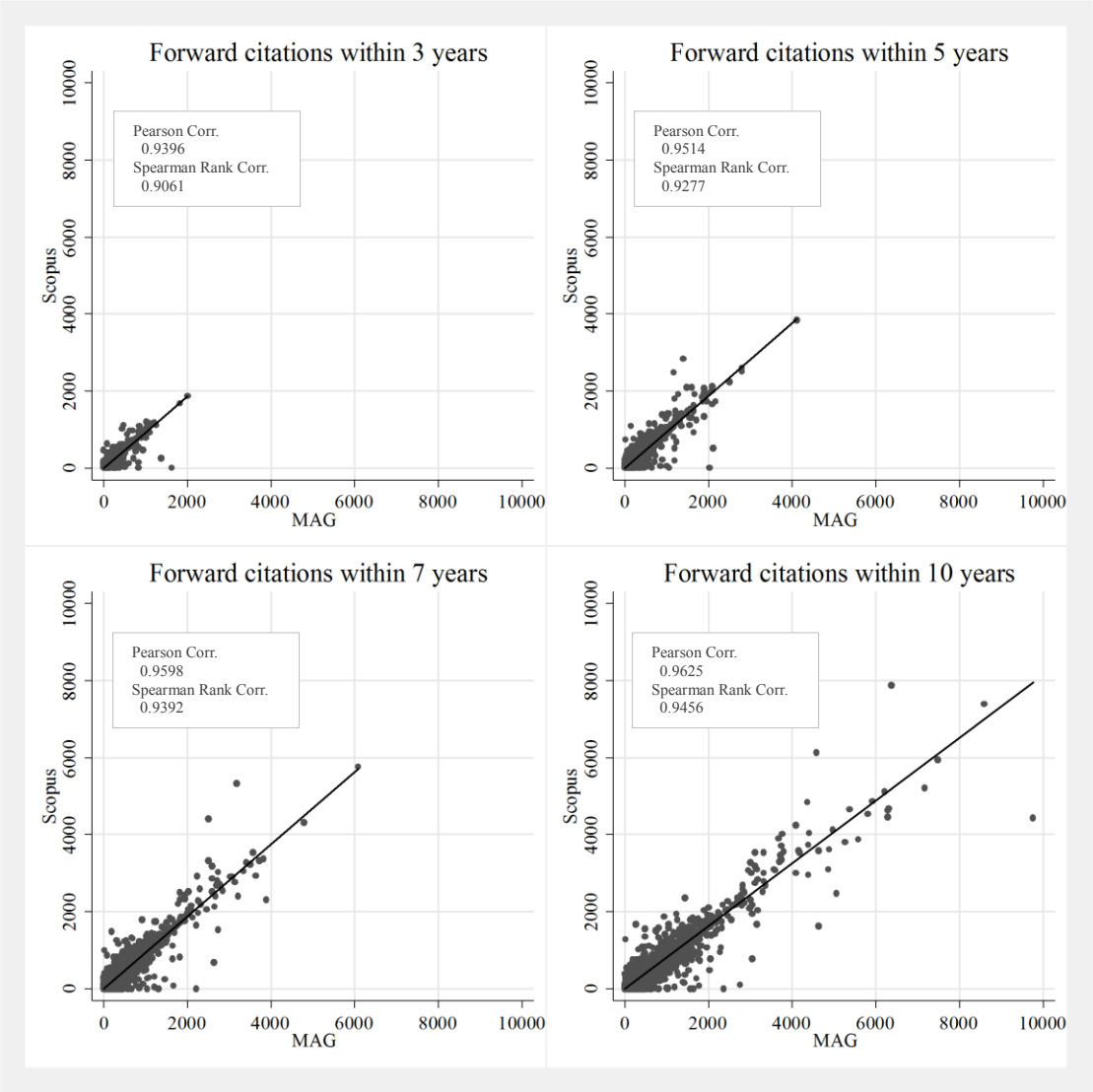


(c) 書誌情報がある後方引用文献：5 年以内の文献のみ



（出典：著者）

図 6. 散布図：前方引用数の比較（2005 年に出版された文献）



（出典：著者）

表 10：MAG の論文毎にみた氏名と所属機関情報の有無

		Affiliation Information			
		Yes	Partly yes	No	Total
Name Information	Yes	43,739,676	2,909,559	119,540,507	166,189,742
	Partly yes	4	3	1,553	1,560
	No	1	0	705	706
	Total	43,739,681	2,909,562	119,542,765	166,192,008

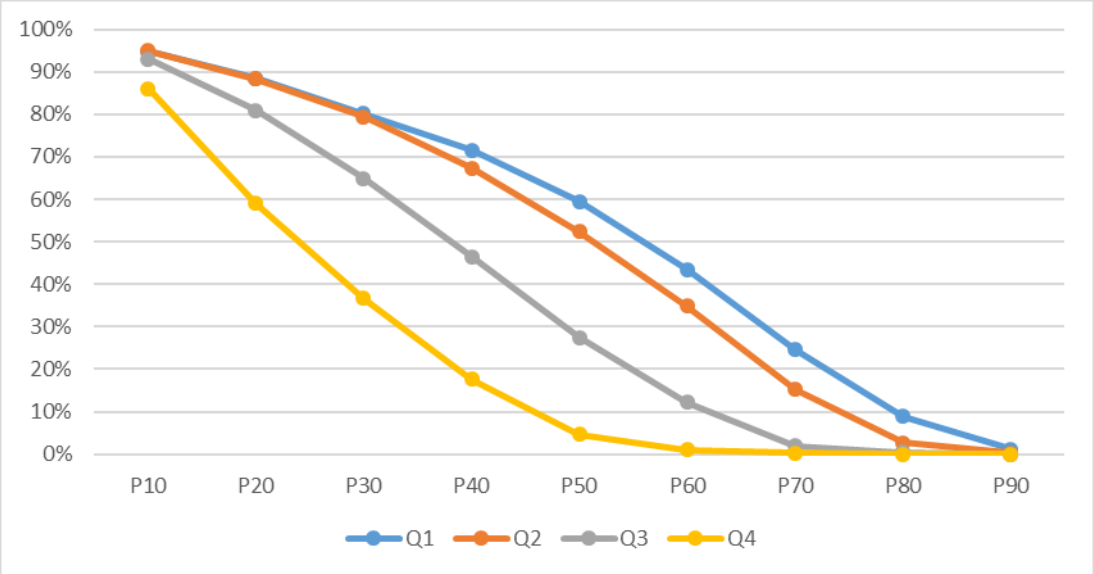
（出典：著者）

表 11：機関情報の有無と国コードの抽出状況（SCIMAGO 論文のみ）

		Affiliation Information			
		Yes	Partly yes	No	Total
Country Code	Yes	21,034,402	0	0	21,034,402
	Partly yes	1,529,423	963,042	0	2,492,465
	No	5,242,599	564,494	25,500,231	31,307,324
	Total	27,806,424	1,527,536	25,500,231	54,834,191

（出典：著者）

図 7：学術誌の質の違いと機関情報付与率分布（SCIMAGO 論文のみ）



（出典：著者）

DISCUSSION PAPER No. 162

Microsoft Academic Graph の書誌情報データとしての評価

2018 年 10 月

文部科学省 科学技術・学術政策研究所 第 1 研究グループ
塚田 尚稔・元橋 一之

〒100-0013 東京都千代田区霞が関 3-2-2 中央合同庁舎第 7 号館 東館 16 階
TEL: 03-3581-2396　FAX: 03-3503-3996

Assessment of Microsoft Academic Graph as a Bibliographic Data Source

October 2018

Naotoshi Tsukada and Kazuyuki Motohashi

First Theory-Oriented Research Group
National Institute of Science and Technology Policy (NISTEP)
Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan

http://doi.org/10.15108/dp162

http://www.nistep.go.jp