STI Horizon 2018 Vol.4 No.4
http://doi.org/10.15108/stih.00158

レポート

Microsoft Academic Graph の書誌情報データベースとしての評価

第 1 研究グループ 主任研究官 塚田 尚稔、客員研究官 元橋 一之

概　要

エビデンスベースの政策形成に資する研究を進める上で書誌情報データベースは重要である。これまでは Clarivate 社の Web of Science（WoS）や Elsevier 社の Scopus が信頼性の高いデータベースとしてよく利用されてきたが、近年は他の選択肢が幾つかでてきた。その一つである Microsoft 社の書誌情報データ Microsoft Academic Graph（MAG）は無償でダウンロードできるバルクデータが公開されたこともあって関心を集めている。科学技術・学術政策研究所（NISTEP）第 1 研究グループでは、このデータベースの利用可能性について Scopus をベンチマークとして大規模サンプルで検証した。結果は既に NISTEP ディスカッション・ペーパー[1] として公表しており、本稿ではその内容を紹介する。

キーワード：書誌情報，データベース，Microsoft Academic Graph，Scopus

1. はじめに

科学技術イノベーションの分野などでは、エビデンスベースの政策形成に資する研究のために、書誌情報を利用した分析が重要視されている。大学や研究機関の研究活動を支える政策の評価や大学ランキングの作成、研究者の評価など、研究活動における先行研究調査以外にも、各国において多岐にわたる用途で利用されている。

経済学に限ると、書誌情報を用いた研究は、学術文献よりもむしろ特許等の産業財産権の書誌情報を用いた研究の方が先行してきた。特許データを用いた分析は 1980 年代ごろから関心を集めるようになり、2000 年ごろから論文数が大きく増加した。研究者の裾野が広がった一つの契機としては、誰にでもアクセスしやすい特許データベースが公開されたことが挙げられるだろう。

米国特許については 2001 年に NBER Patent Citation Data File[注1,2]、日本の特許については 2005 年に IIP パテントデータベース[注2] が無償で公開された[3~5]。欧州特許庁が世界中の特許の書誌情報を収録したデータベース EPO Worldwide Patent Statistical Database（PATSTAT）[注3] を比較的安価に提供するようになってからは、世界的にはこのデータベースを利用している研究者が多いと思われる。PATSTAT の収録情報は、利用者からのフィードバックも取り込んで、徐々に正確かつ詳細なものに進化している。一方で、従来の商用データベースも、もともとの主たる用途である企業等の研究開発における先行文献調査などにおいて健在である。

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。どちらもブラウザや API で手軽にアクセスできるウェブサービスも提供しているが、研究目的では包括的にバルクデータを使って自由

注 1 The National Bureau of Economic Research (NBER), http://www.nber.org/patents/
注 2 一般財団法人知的財産研究教育財団 知的財産研究所 (IIP), http://www.iip.or.jp/patentdb/
注 3 European Patent Office (EPO), https://www.epo.org/searching-for-patents/business/patstat.html

（2018.12.20 公開）

に加工・集計したいというニーズがある。学術論文データベースに関心をもつ研究者は潜在的には多いと思われるが、この分野の研究に気軽に手を出せない理由の一つは、バルクデータが高額であることだろう。大規模なデータを取り扱うスキルをもつ、又は学ぶ意欲と時間のある若手研究者などのためにも、環境の改善が期待される状況である。

Microsoft 社は 2015 年に書誌情報データベース Microsoft Academic Graph（MAG）を公開した[6)]。Google 社が 2004 年から展開している Google Scholar と同様に、MAG はウェブのクローリングによって収集した書誌情報を整理して収録したデータベースである。MAG はバルクデータが無償で公開されたこともあって関心を集めている。

図表 1 には、MAG、WoS、Scopus に収録されている文献数の推移を示した。1980 年から 2016 年の期間において WoS の収録文献数は合計 5,325 万件、Scopus は 5,566 万件であるのに対して、MAG は 1.4 億件で 2 倍以上の件数である。WoS と Scopus は収録するジャーナル等のアイテムを一定の基準で選択しているため一概に比較できないが、MAG の収録文献数は多いといっていいだろう。

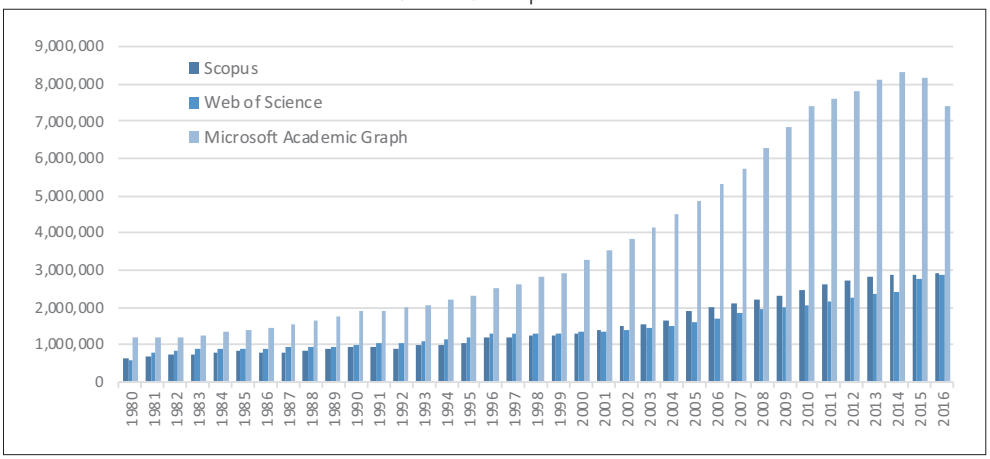
書誌情報を用いた研究においてデータベースの特徴を理解しておくことは重要である。2016 年ごろから MAG の文献のカバレッジや情報の正確性を検証した複数の論文が公表されてきた[7~10)]。しかしながら、これらは特定の機関の論文等に注目したものなどであり、データベース全体の状況を代表する情報を提供しているとはいえない。塚田・元橋（2018）では、MAG の利用可能性について Scopus をベンチマークとして大規模サンプルを用いて評価した。本稿ではその結果の一部を紹介する。

2. Microsoft Academic Graphについて

MAG[注5] のデータにアクセスする方法は幾つかある。Microsoft Academic のサイト[注6] においてキーワードなどで検索をする方法のほかに、Academic Knowledge API[注7] 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。Google Scholar と異なり 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 MAG の有用な特徴である。しかし、今回、我々が用いたのは Open Academic Society のウェブサイト[注8] において提供されているバルクデータである。これは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）。

我々が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

図表 1　MAG, WoS, Scopus の収録文献数[注4]



注 4　Web of Science の件数は、http://apps.webofknowledge.com において Web of Science Core Collection に収録されている文献数を PY= 1980-2016 の条件で 2018 年 10 月 11 日に検索した結果である。Scopus の件数は、https://www.scopus.com において PUBYEAR AFT 1979 AND PUBYEAR BEF 2017 の条件で 2018 年 10 月 12 日に検索した結果である。MAG の件数は第 2 節で説明する 2017 年までのデータを収録しているバルクデータに基づく。MAG の最新データにアクセスするためには後述の API を用いる方法などがある。

注 5　https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

注 6　https://academic.microsoft.com/

注 7　https://labs.cognitive.microsoft.com/en-us/project-academic-knowledge

注 8　https://www.openacademic.ai/

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・システムである MySQL、及び統計分析用ソフトウェアの Stata に読み込んでデータの加工と集計を行った。

MAG の収録データ

MAG には 1800 年以降の 166,192,182 件の文献が収録されている。収録文献数が多くなるのは 1980 年代以降であり、1990 年以降の文献だけで全体の 78%を占める。

MAG に収録されているデータ項目は、図表 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[注9]、著者の所属組織、キーワード、研究分野の分類、参考文献、データソース URL の項目については一つの文献に複数のレコードがある 1 対多の構造になっている。

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、全体の 37% に収録されている。文献タイプ（doc_type）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、ジャーナル掲載論文が中心である。言語の分類（lang）は全体の 85% の文献に情報があり、このうち英語の文献が 8,680 万件で 61% を占める。次いで、日本語の文献が 1,212 万件（8.6％）ある。また、スペイン語 576 万件（4.1%）、中国語 563 万件（4.0%）、フランス語 449 万件（3.1%）、ドイツ語 251 万件（1.8%）などの言語の文献も多い。Field of Study（FOS）は MAG 独自の研究分野の分類データであり、文献のキーワード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、最上位の Level 0 は 19 分類[注10] である。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[11)]。

3. Scopus との比較

3-1. サンプル

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、MAG と Scopus それぞれから抽出した出版年、著者数、後方引用（参考文献）数、前方引用（被引用）数を比較した。

MAG との比較分析のために我々が利用した Scopus データは科学技術・学術政策研究所（NISTEP）が 2014 年度にバルクデータとして購入したもので、主に 1996 年から 2014 年に出版された 34,961,473 件の文献の書誌情報を収録したものである。このうち DOI の情報がある文献は約 60% である。古い文献ほど DOI がない文献が多い。一方、MAG では、1996 年以降では DOI がある文献のシェアは 40% 前後で推移している。MAG と Scopus のレコードが 1 対 1 で接続できた文献数は 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した。1996 年から 2014 年の期間において、このサンプルは Scopus の DOI 付き文献の 91%、MAG の DOI 付き文献の 50% に当たる。

3-2. 比較した結果

出版年と著者数については MAG の情報の精度はかなり高く、出版年は比較分析サンプルの 97.0%、著者数は 98.8% の文献において Scopus の情報と一致した[注11]。個々の文献の著者名の精度検証については今後の課題である。

参考文献（後方引用）の情報については、MAG では比較分析のサンプルの約 15% の文献で欠損[注12] している。オンライン・ジャーナルのウェブページの構造によっては、特定の情報をクローリングで収集できない場合があるのが一つの原因と考えられる。また、MAG と Scopus では参考文献情報の収録方針の違いがあることにも注意が必要である。Scopus の場合は、全ての参考文献に ID を付して、その参考文献 ID

注 9　フルネームの著者名やファーストネームがイニシャルで略された著者名などが混在したデータである。

注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。

注11 情報が一致しない文献については、MAG のクローリングの問題と思われる場合も多いが、MAG の情報が間違っているケースだけとは限らない。例えば、著者数が一致しない文献の中には Scopus の著者名が「et al.」であるようなケースも存在した。情報が一致しない原因については引き続き検証が必要である。

注12 参考文献が全くない学術論文はかなり限定的な数であると思われるため、ここでは参考文献数がゼロである場合は、参考文献情報が欠損しているとみなした。MAG には参考文献情報があるが Scopus にはないケースも 2.1% ほど存在した。なお、質の高いジャーナルの方が MAG の参考文献の収録率が高い傾向がある。

図表 2　MAG 収録データ項目

データ項目	説明	データ収録数	収録率	対応関係
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	研究分野の分類	109,993,272	66.2%	1 : 多
url	データソースのURL	161,847,144	97.4%	1 : 多

を収録している。ただし、参考文献 ID があっても、その書誌情報が Scopus に収録されているとは限らない。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参考文献として収録されている。したがって、参考文献数を正確に知りたいときは Scopus を利用すべきある。参考文献の書誌情報も利用して分析する必要があるならば、MAG の方が使いやすい場合もあると思われる。

MAG と Scopus の参考文献数を具体的にみてみる。両方のデータベースに少なくとも 1 件の参考文献 ID がある文献（全体の 82.8%）に注目すると、平均参考文献数は Scopus が 33.1 で MAG は 27.4 であり Scopus の方が多い。書誌情報とリンクされている参考文献が少なくとも 1 件はある文献（全体の 78.7%）に注目すると、MAG の平均参考文献数は 28.4、Scopus では 19.5 で、Scopus よりも MAG の方が大きな値であった。

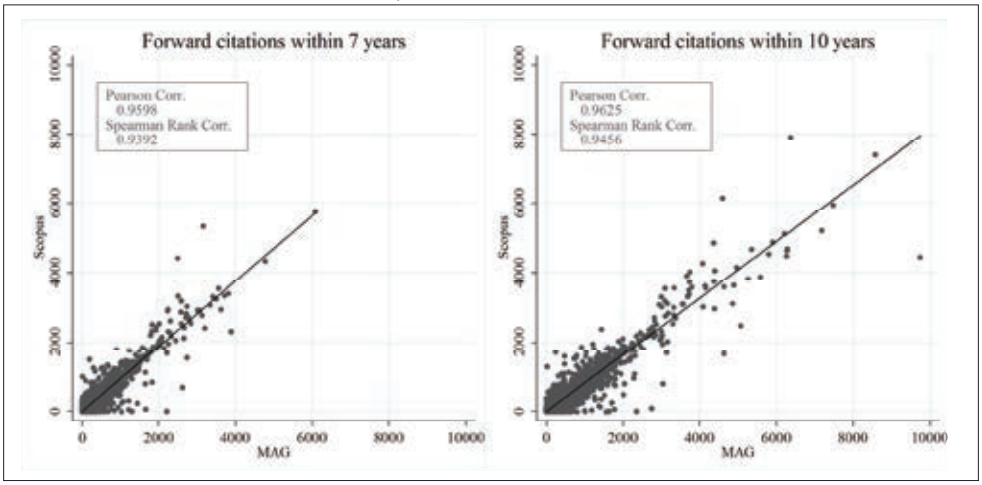
前方引用数（被引用数）は論文の質を測る指標としてよく用いられる。既に述べたように、今回の分析では Scopus は 2014 年度に購入したデータを用いているため、MAG よりも文献収録期間が短い。この違いを考慮して、論文の出版後 7 年以内及び 10 年以内に引用した文献を各データベースでカウントして作成した前方引用数を比較した。

2005 年に出版された文献（965,695 件）について、まず、出版後 7 年以内の前方引用数に注目してみる。この場合は二つのデータベースの文献収録期間の違いによる影響はないと考えられる。図表 3 に示した散布図をみると、対角付近にプロットされた観測値が多く、MAG と Scopus の前方引用数に大きな差はない[注13]。出版後 10 年以内の前方引用数では、収録期間の違いの影響もあるためか、MAG の前方引用数の方が全体的に大きい傾向にある。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 0.963、スピアマンの順位相関係数は 0.946 であった。大規模に集計すると Scopus と MAG は同じような傾向を示す。

ジャーナルごとの参考文献の収録状況の違いを詳細に確認することは今後の課題である。被引用数は収録文献数の多い MAG の方が大きい値になると考えられるが、Scopus の方が大きな値であるケースもある。MAG はウェブのクローリングで書誌情報を収集

注13 2005 年出版の文献について、Scopus から作成した出版後 7 年以内の前方引用数の中央値は 8、平均値は 19.5 であり、MAG の前方引用数の中央値は 8、平均値は 19.7 で、両者はかなり近い値である。もう少し古い文献では Scopus で作成した前方引用数の方が少し大きな値になる傾向がある（2000 年に出版された文献では、Scopus の平均値 19.6、MAG の平均値 17.9、ピアソンの相関係数は 0.965）。

図表 3　2005 年出版文献：Scopus 及び MAG から作成した前方引用数の散布図



している。ウェブページの構造はジャーナルごとに統一されているはずなので、参考文献情報の欠損状況もジャーナルごとに偏っている可能性が高く、MAG と Scopus の被引用数の差に影響を与えているものと考えられる。

4. 論文掲載誌と著者所属機関情報のカバレッジ

論文書誌情報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・ファクターからみた質に関する情報）や論文著者の所属機関に関する情報を整理することが必要である。

論文掲載誌については、MAG ではジャーナル名の情報が Venue の項目に収録されているが、これに ISSN を付すことができれば論文の学術分類や学会誌の学術ランキング情報を容易に接続して利用することが可能となる。塚田・元橋（2018）では、MAG のジャーナル名のテキスト情報に ISSN を付す作業を試みた。論文総数 166,192,182 件のうち、MAG の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 を接続できた。3 節の比較分析のサンプル（19,166,705 件）に注目すると 15,355,987 本（全体の約 8 割）について MAG から ISSN 情報が得られることが分かった。MAG 全体からみると、ISSN 情報を付与できた論文数は 1 / 3 以下となるが、Scopus 収録論文についてみ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。

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。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[注14]。

全ての著者において所属機関情報が存在する文献は約 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 億万件については所属機関情報なしになっている。この点についても、ウェブページ構造とクローリングの問題が原因であると考えられ、情報の欠損状況はジャーナルごとに決まっている可能性が高い。塚田・元橋（2018）では、相対的に質の高いジャーナル論文の方が所属機関情報の収録率が高い傾向があることを報告している。大手出版社の電子ジャーナルの方がメタデータの統一などが進んでいるためと考えられる。現状では、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は不十分だろう。

なお、NISTEP では Scopus-NISTEP 大学・公的機関名辞書対応テーブル[注15] をウェブサイトで公表している。日本の機関に限られるが、Scopus に収録されている著者所属機関名を名寄せした結果と Scopus の文献 ID の対応関係をまとめたものであ

注14 塚田・元橋（2018）では Stanford Named Entity Recognition System（Stanford NER）を用いて機関の所在地である国コードを作成することを試みている。

注15 文部科学省科学技術・学術政策研究所 & エルゼビアジャパン株式会社（2018）『Scopus-NISTEP 大学・公的機関名辞書対応テーブル（ver.2018.1）』 http://www.nistep.go.jp/research/scisip/randd-on-university.

る。出版年が 1996 年から 2016 年の文献単位で 235 万件のデータを収録しており、158 万件の文献には DOI のデータがある。DOI を用いて 152 万件について MAG に接続することができた。これらの文献については、Scopus で著者の所属機関情報を補完して利用することが可能である。

5. まとめ

本稿では、Microsoft 社の書誌情報データベース Microsoft Academic Graph の利用可能性について、Elsevier 社の Scopus をベンチマークとして評価した結果を紹介した。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 と遜色ない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損している論文が多く、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り、ウェブ上に存在する論文等の研究成果を全体的に把握した上で議論する目的には役立つといえるが、機関別の評価など研究目的によっては従来の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だろう。

今回は、Scopus との比較をベースに MAG の評価を行ったが、今後の課題として、まず WoS との比較を挙げることができる。また、MAG の特性について更に検証するためには、ジャーナルごとの分析を進めることも有益である。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

1) 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, National Institute of Science and Technology Policy, Tokyo, DOI: http://doi.org/10.15108/dp162.
2) Hall, B. H., A. B. Jaffe, M. Trajtenberg（2001）"The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools," NBER Working Paper No.8498, DOI: https://doi.org/10.3386/w8498.
3) 後藤晃・元橋一之 (2005)「特許データベースの開発とイノベーション」知財研フォーラム , 63, pp.43-49.
4) Goto, A. and K. Motohashi (2007) "Construction of a Japanese Patent Database and a First Look at Japanese Patenting Activities," Research Policy, Vol.36, Issue 9, pp.1431-1442, DOI: https://doi.org/10.1016/j.respol.2007.06.005.
5) 中村健太 (2016)「『IIP パテントデータベース』の開発と利用」国民経済雑誌, 214 (2), pp.75-90.
6) 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: https://doi.org/10.1145/2740908.2742839.
7) Harzing, A. (2016) "Microsoft Academic (Search): a Phoenix arisen from the ashes?" Scientometrics, Vol.108, Issue 3, pp.1637-1647, DOI: https://doi.org/10.1007/s11192-016-2026-y.
8) Harzing, A. and S. Alakangas (2017a) "Microsoft Academic: is the phoenix getting wings?" Scientometrics, Vol.110, Issue 1, pp.371-383, DOI: https://doi.org/10.1007/s11192-016-2185-x.
9) 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: https://doi.org/10.1007/s11192-017-2454-3.
10) 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: https://doi.org/10.1007/s11192-017-2535-3.
11) Hug, S. E., M. Ochsner and M. P. Brändle (2017) "Citation Analysis with Microsoft Academic," Scientometrics, Vol.111, Issue 1, pp.371-378, DOI: https://doi.org/10.1007/s11192-017-2247-8.