バイオインフォマティクスを使い尽くす秘訣教えます！【第2回】

AllieとColilの使い方
―PubMed/MEDLINEから効率よく情報を抽出する日本発のサービス―

山本　泰智

毎日3000報以上．2015年中に生命科学系の書誌情報データベースとして世界的に広く利用されているPubMed/MEDLINEに新規に登録された件数の平均値である．本稿執筆時点（2016年7月7日）では検索可能な全書誌情報として実に26,222,662件が登録されており，指数関数的に増加している．このような状況においては自身の研究分野だけでも最新の文献情報をすべて網羅的に把握することは困難であると考えられることから，コンピューターを利用したさまざまな検索技術を駆使することが自身の研究を効率的に進めていくためには欠かせない．本稿ではPubMed/MEDLINEの提供する検索機能や筆者の所属するライフサイエンス統合データベースセンター（DBCLS）で開発・提供している2種類の文献情報関連サービス（Allie，Colil）を紹介し，読者の皆さまが的確に必要な文献情報，さらには研究動向を把握するすべを効率よく習得できるようにしたい．

PubMed/MEDINEの歴史と最新情報

上記の通り非常に多くの書誌情報を納め，世界的に広く利用されている書誌情報データベースおよび検索サービスのPubMed/MEDLINE[1]であるが，現在のようにインターネットを介して誰でも自由に無料で検索できる環境が整えられたのは今から20年前の1996年である．それまでは書誌情報データベースとしてMEDLINEが1971年から米国National Library of Medicine（NLM）により旗艦データベースとして維持管理されていた[2]．公開当初のころは検索するために米国内のいくつかの都市に設置されたアクセスポイントに対して有料電話回線経由でのダイアルアップ接続が必要であった．さらに同時アクセスの上限数がわずか25であったほか，検索を行うためには専門知識が必要で，利用に際して各機関につき最低1名がNLMで開催される講習会に参加することを求められていたようだ．その後80年代にはMEDLINE検索用のソフトウェアGrateful Medが開発されたが，検索には費用が発生した．

公開当時のMEDLINEには239の学術誌にて1969年以降発表された13万件超の書誌情報が納められている程度であった．それが2016年では，主に1946年以降に5500を超える学術誌に掲載された論文を対象とした2600万件を超す書誌情報が格納されるに至っている．そのPubMed/MEDLINEに対する検索件数，利用者数，閲覧数は，それぞれ，1日当たり250万件，160万人，そして830万件という状況である．

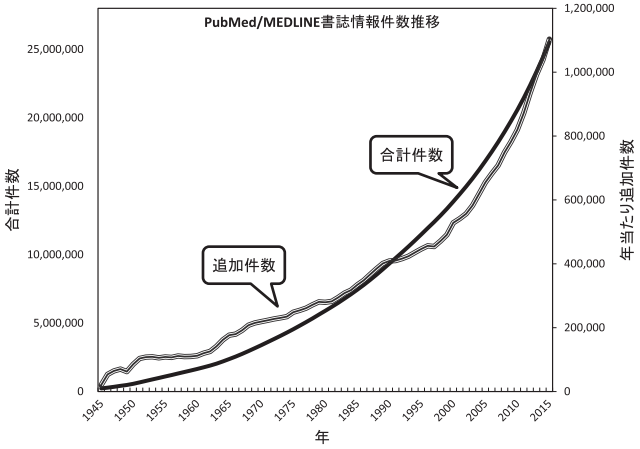

図1．PubMed/MEDLINEの増加

図1は毎年の追加件数と，累積の件数，すなわち検索対象となる書誌情報の総数の推移を表している．研究分野の発展を反映して指数関数的に増加していることが分かる．MEDLINEに対するインターネット経由での検索サービスとしてPubMedが登場した1996年には検索対象となる書誌情報は1200万件弱で前年1995年中に追加された書誌情報は約46万件であった．両者共にこの20年で倍以上になっている．また文献当たりの著者数が年々増加しているという傾向もみられる．

MeSHとPubMed/MEDLINE

毎年指数関数的に増加するPubMed/MEDLINEに納められている書誌情報から所望の論文を見つける際にはMedical Subject Headings（MeSH）タームを適切に用いることが効果的である．MeSHとはNLMにより構築さ

著者紹介　ライフサイエンス統合データベースセンター（特任准教授）　E-mail: yy@dbcls.rois.ac.jp

れた，概念階層をもつ統制語彙（シソーラス）であり，各論文において扱われている研究課題を端的に表す概念としてMeSHタームが5から15語つけられ索引として利用されている．MeSHの概念階層は解剖や生物，疾患といった16のカテゴリーを頂点とし，それぞれに意味的に下位の，すなわち，より狭義な概念が定義されている．概念階層の深さはMeSH全体で一定ではなく，2016年現在でもっとも深い階層は12である．階層の深さと概念の抽象度は絶対的なものではなく，それぞれの階層関係における相対的なものであるほか，上位概念が複数あるものも含まれている．図2はNLMの提供するMeSHブラウザの一ページであり，16のトップカテゴリーが表示されているほか，例として生物（Organisms）の直下にある概念を表示させている．すなわち，生物カテゴリーには真核生物（Eukaryota），古細菌（Archaea），細菌（Bacteria），ウィルス（Viruses），生物形態（Organism Forms）からなり，さらに，各概念を展開していくと次々にそれらの直下の，より狭義な概念を確認できる．たとえば，真核生物であれば，アルベオラータ（Alveolata），アメーバ（Amoebozoa），動物（Animals）などである．

研究の進展を適宜反映できるようにMeSHは毎年改訂されており，納められている概念数は最新のもので約2万8千ある．この中から各論文の内容を表すのに相応しいものを人手で選択しているため発表直後の論文にはMeSHタームが振られていない．担当スタッフはすべて生命科学系の学士以上の学位を有する約100名で，新規に論文が発表されると全文に目を通し，相応しいMeSHタームを選択する[3]．ただ，すべてが手作業ということではなく，論文情報を与えると自動的に候補となるMeSHタームを出力するプログラムの支援を受けて作業がなされている．

各論文に与えられたMeSHタームのうち，特に対象となる論文で主に扱われている概念にあたるものにはアスタリスクマーク（*）がつけられている．また，PubMed/MEDLINEの書誌情報に書かれている各MeSHタームには，続いてさらに詳しい文脈を与えるサブヘッディング，あるいはクオリファイアと呼ばれる語が複数つけられていることがある．たとえば代謝（metabolism）や遺伝学（genetics），単離および精製（isolation & purification）などである．一つのMeSHタームに複数のサブヘッディングがつけられることも多い．たとえば，「Induced Pluripotent Stem Cells/*cytology/metabolism」と表示される場合もあり，これについては，MeSHタームがInduced Pluripotent Stem Cellsで，cytologyおよびmetabolismがサブヘッディングとなる．そして特に細胞学（cytology）に関する論文であることをアスタリスクにより示している．

実際にはPubMed/MEDLINEで興味ある検索語を入力すると，システムが自動的に関連度の高いと計算されるMeSHタームを補完して検索が実行されている．その内容は，検索結果一覧のページで右側の表示部分にある「Search details」という項目名の領域に示される．たとえば，「ips cells」と入力して検索すると，次のように

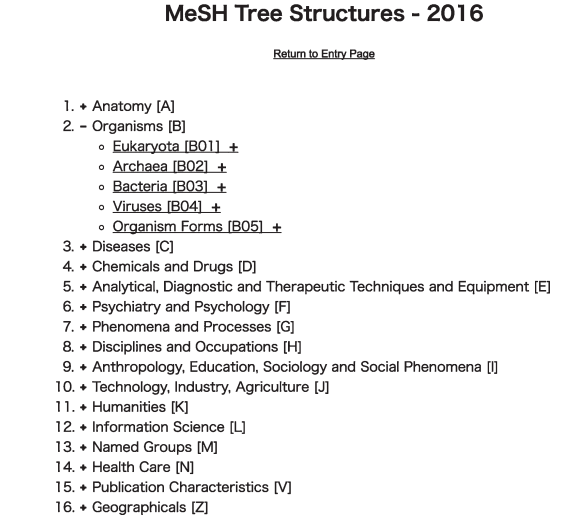

図2．MeSHの概念階層最上位にあたる16のカテゴリーと，その一つである生物の直下にある概念

図3.「ips cells」でPubMed検索した結果（左）とその中から一件を選択してMeSHタームを表示させた結果（右）．検索結果ではPubMedにより自動生成された検索語にマッチした部分が太字で表示されている．

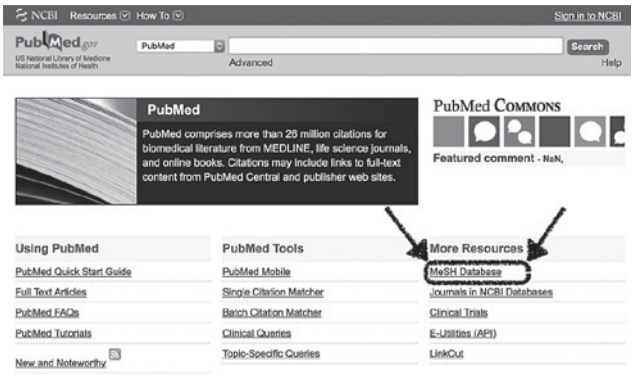

図4. PubMedのホームページからMeSH Databaseへのリンクがある

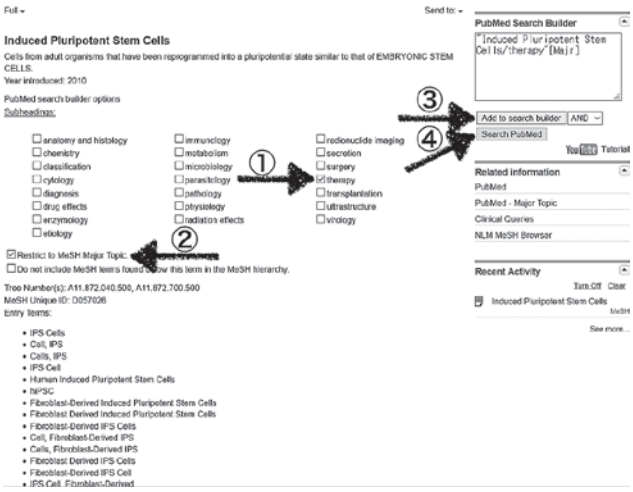

図5. MeSHタームを用いた高度な検索語の構築

非常に長い検索語が自動的に生成されて検索が行われていることを確認できる.

“induced pluripotent stem cells”[MeSH Terms] OR (“induced”[All Fields] AND “pluripotent”[All Fields] AND “stem”[All Fields] AND “cells”[All Fields]) OR “induced pluripotent stem cells”[All Fields] OR (“ips”[All Fields] AND “cells”[All Fields]) OR “ips cells”[All Fields]

ここで角括弧で囲まれているキーワードが，検索方法を指定するためのもので，MeSHタームとしてinduced pluripotent stem cellsが含まれているのがわかる．検索結果を見ると，検索語にマッチする部分が太字になっているほか，各書誌情報に付けられているMeSHタームのリストに検索語として指定されたものが確かに含まれている（図3参照）.

このように，ips cellsと入力しただけでPubMed/MEDLINE側が利用者の意図をくみ取るような仕組みが作られている．ただ，すべてお任せにしていると真に求めている文献を見つけるための検索が必ずしも行われていないこともあり得るわけで，積極的に利用者が上述のサブヘッディングなどを駆使して検索語を生成することもできる．ホームページ（http://pubmed.gov/）にはさまざまな関連データベースへのリンクが表示されているが，そのうち，「More Resources」の「MeSH Database」を選択すると，適切なMeSHタームを探したり，主要な研究課題として指定するMeSHタームを選択したりしながら検索条件の構築と文献検索が行える（図4参照）．なお，検索語の大文字小文字は検索時に考慮されないのでiPS Cellsでもips cellsでも同一の結果が得られる.

上記の検索は，最初に興味のある概念を入力して適切なMeSHタームを選択することから始まる．たとえば先ほどよりさらに短く, ipsとだけ入力してみる. すると，関連するMeSHタームの候補一覧が表示される．そのトップにInduced Pluripotent Stem Cellsがあるが，他にもipsという表現がMeSHタームやその定義に含まれるなど関連すると計算されたものが列挙される．「Induced Pluripotent Stem Cells」を選択すると，その定義や検索時に追加できるオプション群，IPS CellsやhiPSCなどの同義表現，MeSH概念階層中での位置を確認できる．ここでiPS細胞に関連した治療について主に記述している論文を探したいとすると, 図5中①「Subheadings」（サブヘッディング）に「therapy」を選択し, 続いて②「Restrict to MeSH Major Topic」を選択する. そして③右上の「Add to search builder」をクリックすると，その上の領域に対応する検索語が自動的に入力される．これで問題なければ，④「Search PubMed」をクリックする．また，MeSHタームだけでなく，著者名（例：yamanaka[au]）や出版年（例：2014年から2015年に出版された論文に絞る場合は2014:2015[db]）を追加するなど，他の検索オプションを加えたり，検索語を追加したりすることも可能である.

なお，PubMed/MEDLINEの検索システムは上述したようなMeSHタームを用いる方法の他にも実に多くの絞り込み機能を提供しており，すべてを本稿で紹介することはできない．より詳細については参考文献[4]にあたるか，筆者が講習会で発表資料として利用しているスライド[5]を参照していただければ幸いである.

略語の検索サービスAllie

本章以降，DBCLSで開発・提供しているサービスを紹介する．最初はPubMed/MEDLINE中に出現する略語とその意味を示す正式名称（展開形）を簡単に検索できるサービスAllie（アリーと読む）[6]である．使い方は，検索サイト（http://allie.dbcls.jp/）にアクセスすると表示される入力ボックスに関心のある略語を入力して検索を実行するだけである．結果として対応する展開形があればそのリストと関連情報が列挙される．そこから一つを選択すれば出現文献リストや表記のバリエーションなど，より詳細な情報を閲覧できる．

生命科学分野における論文には非常に多くの略語が登場する．自身の研究分野であれば略語の表す意味が分かるだろうが，共同研究の多い昨今では必ずしも自身の研究分野ではない研究者による論文などを読む機会も多くなるだろう．さらに一つの略語が複数の意味を持つ，多義語であることも多い．たとえばPCという略語は，一般的にはpersonal computerの意味で使われていると思われるが，生命科学分野においてはほかにも，primary care，prostate cancerなど多くの意味で使われている．そしてPubMed/MEDLINE中ではphosphatidylcholineの略語としてもっとも多く使われているのだ[注)]．

また，上述の通り多くの論文が日々発表されているが，これに伴い多くの略語もまた生み出されている．我々の調査では月当たり100件前後が新たに論文で紹介されている状況である．このような状況では仮に生命科学分野の略語辞典を編纂するとしても常に最新の略語を含めることは困難である．さらに，後述する機械的な手法により自動的に抽出された，展開形とともに10回以上使われている略語だけでもPubMed/MEDLINE中に72,000種類ほど含まれている．そこで，簡単に最新の略語も含めてその意味や使われている論文などを調べられるサービスが提供されると望ましい．このような需要に応えるべく開発・提供されたサービスがAllieである．

AllieはPubMed/MEDLLINEに書かれている書誌情報のうち，タイトルとアブストラクトを対象として自動的に略語と対応する展開形を抽出し，使われている書誌情報とともにデータベースに格納している．また，ハイフンの有無や大文字小文字の違いなどの表記上の表層的な相違を吸収し，意味的に同じと計算された展開形をグループ化している．これらの処理は全自動で行われており，PubMed/MEDLINEの急速な増加に追随している．

Allieの検索対象は上記の抽出対象書誌情報において略語とその展開形が互いに近傍に出現している場合に限られ，略語のみが出現する場合は抽出されない．このため，たとえばDNAのようにその展開形を併せて書くことの少なくなった略語についてはAllieを用いて検索した場合に得られる書誌情報の数が，当該略語が使われているものすべてと比べて相違が大きくなることもある．

Allieの特徴は4点あり，日本語訳の表示，主な研究分野の表示，出現書誌情報の表示，そして共起略語の表示である．

①日本語訳の表示　Allieは検索対象の展開形に対して人手によりさまざまな辞書を参照しながら日本語対訳をつけている．このため，論文中での出現頻度の高い展開形には英語表記と併せて日本語対訳が表示される．たとえばESと入力すると，「embryonic stem」や「effect size」「electrical stimulation」という順で対応する展開形の候補が表示されるが，それに併せて「胚幹細胞」や「エフェクトサイズ」「電気刺激，電気的刺激」と対応する日本語訳も表示される．また，主な研究分野として「細胞生物学」「スポーツ医学」「神経学，神経内科学」とそれぞれ日本語訳が表示される（図6参照）．

②主な研究分野の表示　略語とその展開形のペアが多く使われる研究領域を書誌情報と併せて表示する．ここで表示される研究領域名は，NLMが各学術誌に対して与えている，それが主に取り扱う研究領域を表すMeSHターム（複数の場合あり）に基づく結果である．現在合計1173のMeSHタームが用いられており，もっとも多くつけられているものは生化学（Biochemistry）

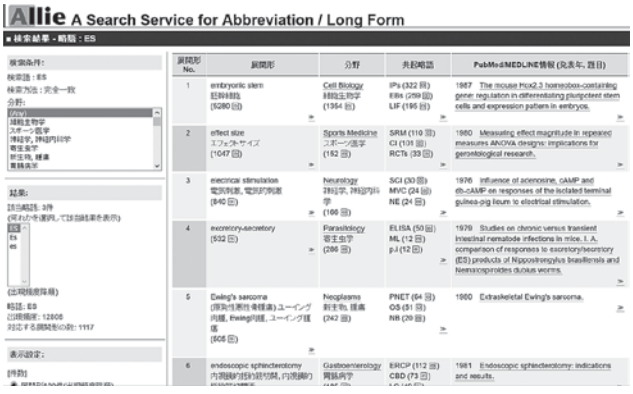

図6．Allieの検索結果例（日本語対訳の表示）

注) 検索結果: http://allie.dbcls.jp/short/exact/Any/PC.html

で，執筆時点（2016年7月19日）におけるAllieでは合計すると303.5万ほどあるペアのうち，12.5万強のペアがこの研究分野に関連している．

③出現書誌情報の表示　Allieでは月一度検索対象データベースを更新しているが，各ペアの出現する書誌情報も更新対象である．このため，常に最新の各ペアの利用状況を確認できるほか，もっとも古い論文，すなわち当該ペアを最初に用いた論文を素早く見つけることができる．

④共起略語の表示　略語の意味を調べようとするとき，それが多義語で複数の展開形の候補が検索結果に出力されると，その中から適切なものを選択する必要がある．Allieでは選択するさいの参考情報として，主な研究分野や書誌情報を表示するが，それにくわえて同じ論文で使われている他の略語も表示する．これは対象の略語の意味が分からなくても共起している略語に自身のなじみの深いものが含まれていれば，それを手掛かりに決めやすいだろうと想定しているからである．

引用情報検索サービスColil

ある論文においてその著者が自身の研究成果を記述するさいには，それまでに得られている関連研究成果との違いを明らかにしたり，研究動機に説得力を持たせたりするなどの理由で他の論文を引用する．この論文の本文中における，引用目的として書かれている部分を引用文脈と呼び，引用される論文を被引用論文と呼ぶ．そして被引用論文ごとに，それを引用する論文（引用論文）の引用文脈を表示するサービスがColil（コリルと読む）[7]である（http://colil.dbcls.jp/）．Colilはさらに，対象となる被引用論文とともに同じ論文から引用される他の論文があればそれらの論文も検索する．これはすなわち，

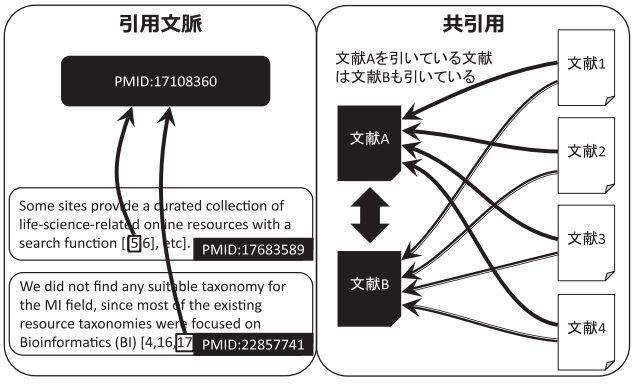

図7．引用文脈と共引用の関係

アマゾンなどのネット通販サイトで採用されている推薦システムと同様に，この論文を引用している論文は次の論文も引用している，という共引用情報を利用者に提示する機能である．図7にColilで用いている語句の関係を示す．

Colilを利用する際にはPMIDが既知の場合はそれを「PubMed IDを入力」と書かれたフォームに入力するが，そうでない場合は，対象となる被引用論文を検索するための検索語を「キーワードを入力」フォームに入力する．Colilは論文を検索するのにPubMed検索を利用しているため，PubMedを利用するのと同一の絞り込み条件を検索語に含められる．たとえば，1995年から2000年の間にJournal of Biological Chemistry（JBC）で発表された細胞死（apoptosis）に関する書誌情報を検索したいとしたときには，「apoptosis 1995:2000[dp] "J Biol Chem"[jour] 」と入力する．そこで表示される検索結果から一件を選択すると，それを被引用論文とする引用文脈が列挙される．併せて共引用情報も提示される．引用文脈が複数ある場合の並び替え方法として，初期設定は発表日時の降順（新しいほど先に表示）であるが，そのほかにも，各引用論文の被引用件数による降順（図8参照）やセクションタイトル順などで表示できる．

Colilを利用することで論文執筆時に関連論文を見つけやすくなるとともに，ある論文から他の論文が引用されている場合には，当該被引用論文を当該論文の著者らがどのように捉えているか効率よく知ることができる．すなわち，論文の著者による要約がアブストラクトであ

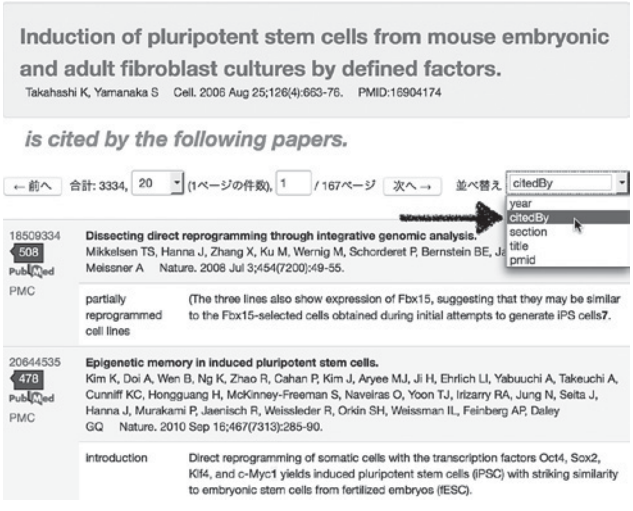

図8．Colilの検索結果を，各引用論文の被引用論文数による降順にする

れば，第三者による要約が引用文脈の集合といえる．この引用文脈の集合は特に英語を母国語としない研究者にとっては，これから自身が引用しようとしている論文に関する英語による記述例としても利用できる．

たとえば，マウスで生成されたiPS細胞に関するCell誌に掲載された論文[8]については，Colilを用いて検索すると，執筆時点で3334件の引用文脈があることがわかる．さらに，当該論文を引用している論文が複数ある場合には，上述の通り，それらの被引用数順に並び替えることができる．そして引用文脈に書かれている内容から，体細胞の再プログラミングがOct4などの4転写因子の形質導入により生じるということ[9,10]や，論文発表後の問題点として，生成されるiPS細胞の質や生成効率が議論されていること[11]などがわかる．続いて共引用論文リストをみると，もっとも併せて引用されている論文は，ヒトの線維芽細胞から生成されたiPS細胞に関するもので，続いてヒトの体細胞からという論文，iPS細胞の生殖細胞生成能力に触れる論文という順で表示されており，iPS細胞に関する研究コミュニティにおいて広く読まれている論文リストが示されていると思われる．

また，Colilにより示される共引用論文の一つを被引用論文として改めて検索するという行為を繰り返すことで，特定の研究分野における主要な論文を効率よく見つけることができる．これは自身が興味を持つ論文がこれまで馴染みのない研究分野である場合に，当該論文が引用している論文の一つをColilで検索することから始めることで，引用情報を基にした，対象研究領域のおおまかな全体像が比較的容易に得られることを意味する．

たとえば，Colilについての論文から引用されている論文に，生物医学論文における参考文献の適切な引用の重要性に関するものがある[12]．当該論文を軸に検索すると，共引用に基づく関連論文として科学的な発表行為における剽窃行為に関する論文[13,14]が見つかり，さらに倫理的な面に焦点を当てた論文[15]が見つかる．このように，生物医学分野での論文発表にまつわる倫理的な側面を扱う研究領域の存在と，そこで議論されている問題点を報告する論文群を比較的効率よくみつけられることに加え，その多くがIzet Masic氏の研究であることがわかるという具合である．

Colilの利用方法について概略は以上の通りであるが，Colilが必要とする情報は主にPMC Open Access（OA）サブセットから取得している．これはCreative Commons（CC）Attributeライセンスなど再利用が自由なオープンライセンスで提供されている論文集合で，National Center for Biotechnology Information（NCBI）が学術誌横断的にすべて同一のXML形式で提供している．このため，論文の本文を含めた全体を対象とした機械的な処理が非常に容易に行える．執筆時点で390万件の論文全文を納めているPMCのうち，PMC OAサブセットは134万件である．なお，PMCではNational Institute of Health（NIH）から取得した予算を用いて得られた研究成果が発表されている論文として著者が全文を提供したもの（NIH Author Manuscripts）についても同形式にて取得可能としており，Colilはこれらからも PMC OAサブセット同様の情報を取得している．この結果，Colilが引用文脈を抽出する対象としている論文，すなわち引用論文の数は，最新版が構築された2016年4月において131万件強であった．また，被引用論文については全文が必要ではなく，PubMedに含まれてさえいれば良いため，同最新版でその件数は814万件であった．PubMed/MEDLINEに含まれる書誌情報の総数を2500万とすれば，それぞれ5%強，1/3強にあたる．

最後に

以上，生命科学分野における文献情報に関するサービスを紹介した．必要な論文を効率よく見つけるだけでなく，論文を執筆する際に役立つサービスもあることを知っていただき，折に触れて有効に活用していただければ幸いである．

文　献

1) PubMed/MEDLINE: http://pubmed.gov/ (2016/7/7)
2) HAPPY 35TH BIRTHDAY, MEDLINE!: https://www.nlm.nih.gov/news/medline_35th_birthday.html (2016/7/7)
3) 阿部信一：情報の科学と技術，**58**, 4 (2008).
4) PubMed Tutorial: https://www.nlm.nih.gov/bsd/disted/pubmedtutorial/cover.html (2016/7/7)
5) AJACS60: http://motdb.dbcls.jp/?AJACS60 (2016/7/30)
6) Yamamoto, Y. *et al.*: *Database*, bar013 (2011).
7) Fujiwara, T. *et al.*: *J. Biomed. Semantics*, **6**, 38 (2015).
8) Takahashi, K., Yamanaka, S.: *Cell*, **25**, 126 (2006).
9) Kim, K. *et al.*: *Nature*, **16**, 467 (2010).
10) Zhu, J. *et al.*: *Annu. Rev. Immunol.*, **28** (2010).
11) Ieda, M. *et al.*: *Cell*, **6**, 142 (2010).
12) Masic, I.: *Acta Inform. Med.*, **21**, 3 (2013).
13) Masic, I.: *Acta Inform. Med.*, **20**, 4 (2012).
14) Armstrong, J. D. 2nd.: *AJR Am. J. Roentgenol.*, **161**, 3 (1993).
15) Masic, I.: *Acta Inform. Med.*, **20**, 3 (2012).