とあるアマチュアの文献索引（インデックス）

中島 淳志

最近はあまりに手軽にできてしまうので実感が湧かないかもしれないが、「検索」は「技能」である。スポーツや芸術と同じで、間違いなく巧拙の差がつく類のものである。菌類に関しても、例えば「2010 年～2015 年に中国から発表された新種が知りたい」「ガラパゴス諸島に分布する菌根菌が知りたい」「シカの糞に発生する子嚢菌が知りたい」…などの様々な具体的疑問（それは研究に関連するものかもしれないし、単純な興味の産物かもしれない）がある時に、答えを含む情報源に的確かつ迅速に辿り着けるかどうかの個人差は大きいと思われる。つまり、簡単ではないのである。特に一般向けの情報の絶対量が少ない菌類の分野では、Google 検索のみでは手に余ることがしばしばある。

先に挙げたような疑問をズバリ掲載しているWeb サイトや一般向けの書籍などはないと思われるので、答えに辿り着くためには学術論文の情報にアクセスする必要がある。しかし、そこには大きな壁が立ちはだかる――いわば、入手の壁、言語の壁、専門性の壁である。幸いなことに最近はオープンアクセスの雑誌も多く、著者自身によるオンラインでの論文の公開も一般的になりつつあるので、入手の壁はかつてないほど低くなっている。そのため、主に問題となるのは目的の文献の特定である。その過程には必然的に外国語と専門知識が関連してくるので、正攻法では高度な技能が要求される。しかし、本稿で紹介する強力な「道具」である「インデックス（索引）」を活用すれば、求める答えを遥かに容易に得られることも多い。検索がスポーツや芸術と異なるのは、よい「道具」があれば、初心者でも相当の――場合によってはプロを凌ぐほどの――アドバンテージが得られるという点である。

文献データベースとインデックス

筆者の職業に関連する医薬分野では、文献情報は高度に体系化されており、例えば医者が目の前の患者に必要な治療を講じる上で、関連する根拠資料（エビデンス）に容易にアクセスできる環境が整えられている。そこで大きな役割を果たしているのは文献情報のデータベースであり、エルゼビア社の「EMBASE」や米国国立医学図書館の「MEDLINE」などが代表例である。これらのデータベースには、文献の著者、出版年、雑誌名といった書誌情報が含まれているほか、文献の内容から抽出したキーワードが「インデックス」として収載されており、データの検索性を高める役割を果たしている。ニコニコ動画やPixiv、あるいはWordPress などに馴染みのある読者には、インデックスではなく「タグ」と言った方が分かりやすいかもしれない。ちなみに、インデックスを付与する作業を「インデクシング」、それを行う人を「インデクサー」という。MEDLINE では関連分野の学士以上の学位を有し、MeSH と呼ばれ

る専用の医学用語集に精通した約100名のインデクサーが作業に従事している[1]。

例えば、ペニシリンを何らかの動物に投与した論文を網羅的に調べたいという時に、「ペニシリン」のキーワードでヒットする情報の中にはまず間違いなく大量の無関係な情報（ノイズ）が含まれている。Google や Yahoo!などの一般の検索エンジンでは、場合によっては同名のロックバンドの情報ばかり出てくるかもしれない。しかし、例えば MEDLINE で「Penicillins/therapeutic use」の MeSH キーワードを使って検索すれば、ヒットする情報はほぼ確実に、ペニシリンの投与に関連する文献情報である。インデックスの威力を理解いただけただろうか。

また、インデックスは検索だけでなく、情報の分類や解析の上でも有用である。医薬分野で過去の研究の知見をまとめた総説記事が盛んに発表されていること、さらには文献情報を高度に統合した「メタアナリシス（メタ解析）」と呼ばれる研究がなされていることには、文献データベースとそのインデックスの存在が密接に関連している。驚くべきことに、相当な割合の医薬論文が少なくとも研究の一部でインデックス（およびそれを基にした検索式など）に依拠しているのである。

では、菌類学分野はどうかというと、医薬分野に比べて文献情報の体系化が大きく遅れていると言わざるを得ない。生物学領域全体をカバーした「BIOSIS」というデータベース（トムソン・ロイター社）はあるが、有料のデータベースであり敷居が高い。MEDLINE のデータの一部は「PubMed」という Web サイトで無償提供されているが、2016年1月現在、菌類分類学の論文を扱う主要な雑誌としては「Mycologia」など少数が収録されているに留まり、大部分の文献は検索できないのが現状である。また、後述するが、広範な分野を扱うデータベースほど、インデックスの種類が個々の分野における利便性や需要に合致しない傾向がある。

表 1 同一論文 (Nagy et al., 2012) に対する MEDLINE、BIOSIS、大菌輪のインデックスの比較
※MEDLINE および BIOSIS のインデックスの和訳は筆者が仮に付したものである。

	MEDLINE	BIOSIS	大菌輪
地理		オランダ/Netherlands	エーランド地方/Öland オランダ/Netherlands スウェーデン/Sweden スロバキア/Slovakia チョングラード県/Csongrád トルナ県/Tolna ヌール・トロンデラーグ県/Nord-Trøndelag ノルウェー/Norway ハンガリー/Hungary バンスカー・ビストリツァ県/Banská Bystrica バーチ・キシュクン県/Bács-Kiskun 南ホラント州/South Holland
宿主・基質			イネ科/Poaceae ウサギ科/Leporidae ウシンケイ属/Festuca ウシ属/Bos ウシ科/Bovidae ウマ属/Equus ウマ科/Equidae カヤツリグサ科/Cyperaceae キク科/Asteraceae シカ科/Cervidae スゲ属/Carex トウダイグサ属/Euphorbia トウダイグサ科/Euphorbiaceae ハネガヤ属/Stipa ヤナギ科/Salicaceae ヤマナラシ属/Populus ヨモギ属/Artemisia
生息環境			ウッドチップ生息菌/wood chip inhabiting fungus リター菌/litter fungi 森林/forest 砂丘/dune 砂地/sand 砂質土壌/sandy soil 芝生/lawn 草地/grassland 粘土質土壌/clayey soil 糞生菌/coprophilous fungi
分子系統解析関連	ハラタケ目[遺伝学]Agaricales/genetics 塩基配列/Base Sequence 菌類のDNA/DNA, Fungal 菌類のリボソームスペーサー領域DNA/DNA, Ribosomal Spacer, Fungal 分子配列データ/Molecular Sequence Data 系統遺伝学/Phylogeny DNAの配列解析/Sequence Analysis, DNA チューブリン[遺伝学]/Tubulin/genetics	配列データ/Sequence data 核リボソームDNA/Nuclear rDNA モデル/Model 推論/Inference	分子系統解析/molecular phylogenetic analysis 祖先形質推定/ancestral character estimation TreeBASE
分類関連	ハラタケ目[分類]/Agaricales/*classification	ハラタケ目/Agaricales *Coprinellus* 属*Setulosi* 節 */Coprinellus sect. Setulosi* ナヨタケ属/*Psathyrella* 分類群/Taxa	タイプスタディ/type study 共有派生形質/synapomorphy 収蔵標本/stored specimen 新種記載/new species 未記載種/undescribed species 検索表/identification key 稀産種/rare species 種の境界/species delimitation 種内変異/intraspecific variation 隠蔽種/cryptic species
その他	ハラタケ目[超微細構造]Agaricales/ultrastructure 菌類の胞子[細胞学]/Spores, Fungal/cytology 菌類の胞子[超微細構造]/Spores, ultrastructure		液化/deliquescence
合計索引数	12	9	53

筆者が運営する非営利私撰データベース「大菌輪」では「論文3行まとめ」と称して、主に 2010 年以降の菌類分類学論文の簡易な抄録情報を提供しており、データの作成にあたって一定のルールに基づきインデックスを付与している[2]。例えば、Nagy et al. (2012) [3]のヒトヨタケ類の系統分類の論文について、MEDLINE、BIOSIS および大菌輪のインデックスを比較すると表1の通りである。

MEDLINE とBIOSIS のインデックスが分子系統解析の関連用語に集中しているのに対して、大菌輪では地理、生息環境、宿主/基質、分類学用語なども索引対象に含めている。この論文には付与していないが、他に実験手法

（例：表面殺菌法、クローニング、HPLC）、生理・生化学（例：温度適性、二次代謝産物、マイコトキシン）、産業・文化（例：生物農薬、きのこ栽培、菌類民族学）などに関連するインデックスも必要に応じて設けている。

2016 年 1 月 18 日時点で「論文 3 行まとめ」で取り上げた 2,569 論文に付与されたインデックスは 4,937 種類、56,342 件であった。カテゴリ別の集計結果を表 2 に示す。大菌輪のインデックスは MEDLINE や BIOSIS のような専門家の手による「商品」ではなく、アマチュアの個人がボランティアで付与しているもので、あくまで参考程度に捉えていただきたいが、このような分野に特化したインデックスは汎用のそれよりも扱いやすく、実際に役に立つ場面も多いのではないかと考えている。

表 2 大菌輪のカテゴリ別タグ一覧

宿主/基質	1896
地理	1329
雑誌	344
生理/生化学	254
実験器具/実験手法/試薬	203
病害/疾患またはその関連事項	168
生息環境	161
分類	135
新産種	132
生態/多様性/保全	130
系統/遺伝/進化	51
産業/食用/文化	40
形態形質	34
データベース	14
化石菌類	13
メタ	13
発表年	13
言語	7

共起インデックスの解析

前述した通り、インデックスは検索だけでなく解析に用いることもできる。例えば、「外生菌根菌」というインデックスが付与された文献に「ブナ科」というインデックスが頻繁に付与されていたら、ブナ科植物と外生菌根を形成する菌が多いのではないか、と推測できる。ある情報（この場合は文献データ）において複数の情報（この場合はインデックス）が同時に出現することを「共起」という。2,569 論文のうち「外生菌根菌」と「ブナ科」が共起する文献は 88 件、「外生菌根菌」のみが出現するのは 171 件、「ブナ科」のみが出現するのは 232 件、どちらも出現しないのは 2078 件であった。以上の「ありあり-ありなし-なしあり-なしなし」のデータが揃えば、以下の数式で「オッズ比」という指標を算出することができる。

$$(\text{オッズ比}) = \frac{(\text{ありあり}) \times (\text{なしなし})}{(\text{ありなし}) \times (\text{なしあり})}$$

オッズ比はごく簡単に計算可能な指標であるが、医薬分野での重要性は高く、例えばデータベースからある「薬」と「副作用」の未知の強い関連（シグナル）を検出するためにも用いられている。オッズ比は 1 から離れるほど強い関連を示すが、シグナルと見なすかどうかには「95%信頼区間（95%CI）」が 1 をまたがないこと（＝5%の危険率で統計学的に有意な差がある）が一つの基準となる。95%信頼区間の上限と下限は以下のように算出される。

$$(\text{オッズ比}95\%CI\text{上限/下限}) = \exp\left(\log(\text{オッズ比}) \pm 1.96\sqrt{\frac{1}{(\text{ありあり})} + \frac{1}{(\text{ありなし})} + \frac{1}{(\text{なしあり})} + \frac{1}{(\text{なしなし})}}\right)$$

先の「ブナ科-外生菌根菌」のオッズ比は 4.60、95%CI 下限は 3.45（>1）なので、両者の関連は強いと判断することができる。これにマツ科など他の 5 科を加えて解析結果を図示したものが図 1 である。このような図をフォレストプロットといい、横棒が縦棒と重なっているかどうかを見ることで容易に有意性を知ることができる。

今回の例は、ブナ科植物の樹下に外生菌根菌がよく発生することを知っていれば、あえてオッズ比を算出せずとも関連性が予測できるが、「薬」と「副作用」のような未知の関連性を、菌類の分野でもオッズ比の適用によって浮き彫りにすることができる可能性

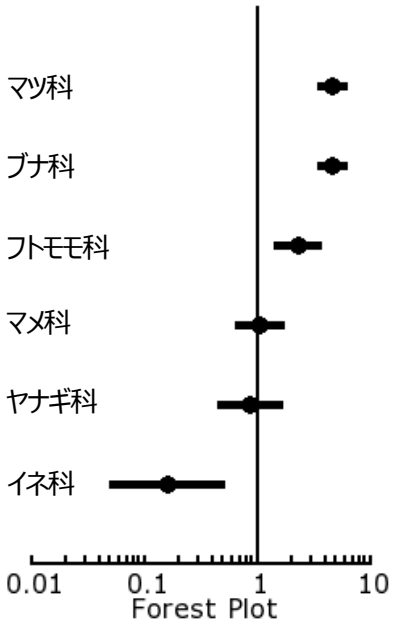

図 1 主な植物の「科」と「外生菌根菌」のインデックスの関連（横軸はオッズ比）

比較対象タグor分類群②	①あ	①あ	①な	①なし	オッズ比	比較対象合計
Agaricales/ハラタケ目	41	166	381	1981	1.28	422
Lecanorales/チャシブゴケ目	0	207	163	2199	0.00	163
Hypocreales/ニクザキン目	2	205	140	2222	0.15	142
Pleosporales/プレオスポラ目	4	203	134	2228	0.33	138
Polyporales/タマチョレイタケ目	81	126	57	2305	**26.00**	138
Pezizales/チャワンタケ目	5	202	90	2272	0.62	95
Boletales/イグチ目	9	198	84	2278	1.23	93
Russulales/ベニタケ目	17	190	69	2293	**2.97**	86
Xylariales/クロサイワイタケ目	9	198	73	2289	1.43	82
Capnodiales/カプノディウム目	1	206	79	2283	0.14	80
Hymenochaetales/タバコウロコタケ目	44	163	33	2329	**19.05**	77

図 2 「菌言マイニングツール」スクリーンショット（上）入力画面、（下）結果出力画面

がある。例を挙げると、「ハラタケ目」と共起する地理関連のインデックスを網羅的に調査したところ、「インド」の関連性が際立って高く、この分類群がインドで近年盛んに研究されていることが示唆された（オッズ比 2.03、95%CI 1.49-2.77）。一方、中国からは近年大量の新種が発表されているが、ハラタケ目との関連は意外にも低く（オッズ比 0.62、95%CI 0.47-0.86）、他の菌類の研究が相対的に盛んであることが示唆された。例えばこれが、「では中国ではどんな菌の研究が盛んなのだろう？」という新たな疑問に繋がるのである[4]。このように、蓄積されたデータから未知の有用な情報を見つけ出す手法は、鉱山での採掘になぞらえて「データマイニング」と呼ばれている。

大菌輪では「論文 3 行まとめ」のデータから様々な条件でオッズ比を算出できるエクセルマクロ、「菌言マイニングツール」を配布している（図 2）。当然ながら筆者も全てのインデックスの組み合わせを試しているわけではないので、ぜひ各々の興味に合わせて未知の価値ある知見を探ってみてほしい。また、オッズ比の算出に必要なのは「ありあり-ありなし-なしあり-なしなし」の 4 つの数字だけであることから、応用の幅が非常に広いことが想像いただけるだろう。例えば PubMed の WebAPI を使って Mycologia の論文のインデックス一覧を取得し、オッズ比を算出して菌類学とその他の学問分野を比較する、といったアイデアも考えられる。

オートインデクシングの未来

もし筆者に時間が無限にあるなら、過去 200 年以上に遡る全ての菌類学の文献にインデクシングを施したいものである。もちろん不可能である。大菌輪のインデクシングは入力を含めて全て人手で行っているが（英訳のみ機械的に照合）、この方法ではあまりにも効率が良くないことを実感している。ただ、機械に全てを任せられるかというと、決して完璧は望めないのが現状である。例えばある文献が「Poaceae（イネ科）」という単語を含

むとしても、大菌輪のインデクシング対象になるのは「掲載されている菌の宿主がイネ科植物である時」に限られる。別のインデックスにはまた異なる採択基準があり、全てを機械に任せるのは現実的でない。この問題を解決するには2つの選択肢があると考えている。まず一つが、あくまで機械による正確な検出を追求する方法である。条件付き確率場（CRF）のような自然言語処理の高度な手法を用いれば、同じ単語でも前後の文脈を基に採否を判定することができるという。しかし、筆者は現在のところ、この手法を使いこなせるだけの技能を持ち合わせていない。もう一つが、機械による判定を諦めるか、あるいは参考程度と位置付け、最低限「人間が判断するのに必要な情報」を余すところなく抽出させる、という方法である。この方法であれば辞書にある単語を網羅的に検出すればよいので、その精度を高めるためには適切な類義語辞書（シソーラス）の構築に注力すれば済むだろう。

おわりに

菌類学の文献情報は宝の山であるが、まさしく深山幽谷であり、訪れる人は稀である。ある分類群の専門家であれば過去から現在に至るまでの全文献を網羅的に把握していることはありうるが、日々世界中から新たな論文が発表されている中で、専門外の分類群、さらには菌類全体ともなると、全貌の理解には到底及ぶことがないだろう。しかし、インデックスの導入によってアクセス性が改善すれば、状況は徐々に変化すると考えている。菌類の公的データベースなどで学術的な使用に堪える質のインデックスが付与されるようになれば、医薬分野のような過去の文献情報をメタ的に統合した研究も増えるかもしれない。

読者の中には、検索云々以前に、「学術論文は専門家のための高尚なものであるから、自分には到底理解できない」という心理的障壁を持っている方もおられるかもしれない。だが、我々が慣れ親しんでいるきのこ図鑑の内容も、元を辿ると、学術論文の形で発表された知見が一般向けに再構成されたものであることが多いので、論文の内容は決して馴染みのないものではない。図鑑やフィールドなどから得た知識を学術論文という強固なエビデンスに結び付けることができれば、その知識はより盤石なものとなり、その過程で豊富な周辺情報や最先端の知見に触れることもできるだろう。学術論文はアマチュアの菌類愛好家（マイコフィル）にとっても大いに有用な資源である。そして、インデックスはその深遠な世界へ一歩踏み出す上での頼れる道しるべとなってくれるだろう。

脚注

[1] 阿部信一（2008）「MEDLINE/PubMed の索引と検索」情報の科学と技術 58(4), 172-177.

[2] 大菌輪ではインデックスを「タグ」と呼んでいる。

[3] Nagy LG et al. (2012) Phylogeny and species delimitation in the genus *Coprinellus* with special emphasis on the haired species. Mycologia. 104(1), 254-275.

[4] ちなみに「中国」と特に関連性が高い「目」は「タマチョレイタケ目」であった（オッズ比 2.32、95%CI 1.57-3.43）