菌類の形質を掌握せよ　～形質データの構造化・統制・集積～

中島 淳志

情報技術の発達に伴い、菌類分類学へのデータサイエンス的手法の適用可能性も広がりつつある。菌類関連のデータベースとしては、学名や分類階級は MycoBank や Index fungorum、出現記録（標本・観察データ）は GBIF、宿主・基質は USDA の Fungus-Host Distribution Database、塩基配列データは GenBank 等が代表的である。しかし、これらのデータベースでは、例えば「傘が黄色のきのこ」「胞子にトゲのあるカビ」等、形質に基づく検索機能は提供されていない。もしそれが可能になれば、分類や同定に有用であるだけでなく、可視化や解析の幅も広がるだろう。しかしながら、形質データは世の中に豊富に存在するにもかかわらず、様々な書籍や論文等に非構造化データとして散在しており、そのままの形では利用困難である。この問題に対処するためには、形質データをどのような手段で収集し、どのような形式で集積し、どのような理念で統制するかを綿密に検討する必要がある。大量の非構造化データを扱うためには、テキストマイニング等の自然言語処理を取り入れて効率的なデータ収集を行うとともに、人間よりもむしろ機械にとっての可読性を重視した（セマンティックな）データベースを構築することが必須であるが、その構築には菌類学の知識と基盤技術の両輪が不可欠である。本稿では、菌類の形質データを誰もが自由自在に取り扱えるようになるという究極的な目標を見据え、その実現手段となりうるデータベース、プロジェクト、および Web アプリケーションを計３点紹介する。

１．形質データベース TraitBank

菌類学専門の形質データベースは本稿執筆時点で、未だ発展途上（0.0.3 版）の funfun database（https://github.com/traitecoevo/fungaltraits/releases）などごく僅かな例を除いて存在せず、構造化データの集積はほぼ全く進んでいないといえるが、他の生物ではその重要性が既に理解されており（分類学ではなく、群集生態学分野における機能形質への着目の影響が強いと思われるが）、全生物のデータを統合する試みも始まっている。TraitBank（トレイトバンク）はその代表例であり、Encyclopedia of Life（EOL）プロジェクトの一部として、2014 年から菌類を含むあらゆる生物（Tree of Life, ToL）の形質（trait）データを集積している。ここで扱う形質には、体長や体色のような形態形質だけではなく、地理的分布、発生時期、宿主、他の生物との相互作用、果てはゲノムサイズ等のあらゆる形質が含まれる。データは Darwin Core という標準形式およびその拡張形式に則って記述され、他の EOL のコンテンツと同様にオープンデータとして自由に再利用することができる。TraitBank のデータは EOL の各分類群のページに一覧表示され、分類群と形質のカテゴリ（属性）の指定によ

る検索画面も提供されている（図 1）。

本データベースの利用可能性を評価するにあたり、まずは菌類形質データの収載状況の把握を試みた。TraitBank の系統的検索の手段としてはWebAPI が提供されているが、グラフデータベースの構造をとっているため、「Cypher QL」という特殊なクエリ言語で検索する必要がある。今回は平易な Web スクレイピングの手法を採用した。

2018 年 12 月 31 日から翌年 1 月 2 日にかけて、MycoBank 収載の全ての合法名（Legitimate）かつ現行学名（Current Name）の 76,022 種（変形菌、卵菌等を含む）で順次 EOL ID の特定および TraitBank の検索を実施し、属性（attribute）、値（value）、および情報源（source）情報を取得した。同種同属性の 6 番目以降のデータは画面上省略されているので取得しなかった。

76,022 種中 23,035 種に EOL ID および 1 つ以上の形質データが存在した。1 種あたりの属性種類数は平均 1.88 種類、最大 7 種類であった。最多の属性は地理的分布（geographic distribution includes, n=21,606）であり、生物間相互作用（interacts with, n=5,970）がそれに続いた（表 1）。ほとんどの属性が生態関連であり、形態関連の形質データは、地衣類の地衣体の形状（樹枝状、痂状、鱗片状等）が僅かに収載されている以外は存在しなかった。一般的な菌類のデータベースであまり見られない属性としては、毒成分（has toxin）や侵略的外来種の移入先（invasive in）等があった。独特な属性としては、線虫捕捉菌の捕食対象（preys on）や水生菌の生息環境（water depth, water O2 saturation, water temperature, etc...）等があった。多くのデータが GBIF のような汎 ToL 的なデータベースに由来し、菌類特有の属性は見出されなかった。横断検索可能な情報源（source）は 226 種類であったが、「parasite of」と「has host」等、ごく一部の属性で概念の重複と見られる例があったほかは、データベース間の属性の統制に問題は認められなかった。

以上の網羅的検索により、菌類の TraitBank データとしては、生態学的形質は比較的充実している一方、形態形質データは未だ著しく不十分であることがあらためて明らかになった。現段階では TraitBank の同定における利用は困難である。汎 ToL 的な基盤が着々と整備されつつあり、哺乳類や鳥類、昆虫類等では分類群によって偏りはあるものの、既に多彩な形態形質が収載されていることを鑑みると、

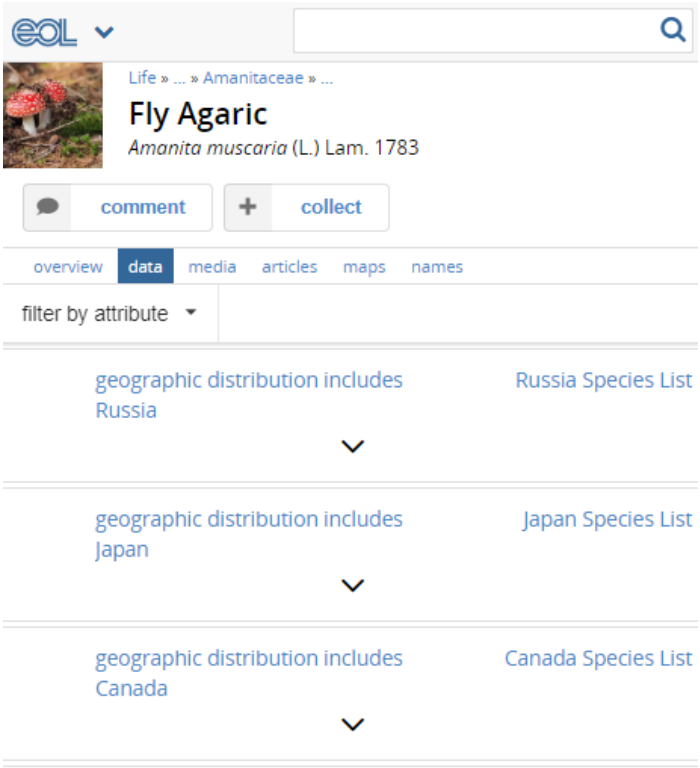

図 1. EOL のベニテングタケのページ
「data」タブから TraitBank のデータ一覧を確認することができる。

表 1. 菌類の種に付与されている TraitBank の属性（attribute）データを集計した表

属性	種数
geographic distribution includes（地理的分布）	21606
interacts with（生物間相互作用）	5970
eats（食性）	2877
habitat includes（生息環境※1）	1763
native range includes（天然分布範囲）	1347
primary diet（主食）	955
parasite of（宿主）	887
growth habit（生長習性※2）	747
symbiotically interacts with（共生関係にある生物）	679
pathogen of（病原体の宿主）	302
feeding method（摂食様式）	209
latitude（緯度）	119
longitude（経度）	119
introduced range includes（移入先範囲）	87
toxic effect（毒性※3）	66
has toxin（毒成分※4）	64
water depth（水深）	45
invasive in（侵略的外来種の移入先）	44
has host（宿主※5）	37
not invasive in（侵略的外来種の非移入先）	23
preys on（捕食対象）	20
conservation status（保全状況）	16
habitat is（生息環境※6）	16
population trend（個体群動向※7）	16
is vector for（媒介者）	14
water silicate concentration（水中珪酸塩濃度）	12
water salinity（水中塩分濃度）	12
water O2 saturation（水中酸素飽和度）	12
water temperature（水温）	12
water phosphate concentration（水中リン酸塩濃度）	12
water dissolved O2 concentration（水中溶存酸素濃度）	12
water nitrate concentration（水中硝酸塩濃度）	12
trophic level（栄養段階※8）	9
extinction status（絶滅状況）	5

※1 森林、沿岸域、乾燥地域、底泥、土壌、市街地、病院…等
※2 地衣類の地衣体の形状（樹枝状、痂状、鱗片状等）
※3 嘔吐、昏睡、発汗、痙攣、眩暈、死亡…等、具体的な症状
※4 イボテン酸。ムスカリン、シロシビン、レンチナン…等、毒成分の種類
※5 "parasite of"との使い分けは不明瞭
※6 地上生、淡水生等、"habitat includes"よりも広い概念
※7 減少傾向にあれば"Decreasing"
※8 寄生菌であれば"Parasite"だが一部の海藻寄生菌のみで使われている

菌類のデータが欠乏しているという現況は、生物多様性研究全体の足並みを乱しかねない危機的状況とえるだろう。研究機関や専門家の主導のもと、速やかにデータ拡充に取りかかるべきだと考える。

２．日本語図鑑・論文情報を情報源とする形質データベース構築プロジェクト

神奈川キノコの会では現在、日本語の各種きのこ図鑑および論文を基に形質データベースを構築するプロジェクトを進めている。このデータベースを野外勉強会での同定支援に利用するのが当面の目標である。また、菌類の形質そのものについて理解を深めること、本会が多数抱える「仮称」の実体を明確化した上で発信することなども視野に入れている。

本データベースに含まれるデータは、「CCC（統制形質組み合わせ，Controlled Combination of Characters)」を単位としており、これは記載文の形質を辞書ベースで自動的に一定の統制語にまとめたものである。例えば「傘が黄褐色」という記述があった場合、「＜傘の色 色=“黄”＞」「＜傘の色 色=“褐”＞」という２つの CCC を生成する。例えば「レモン色」や「淡黄色」と書かれていても同じ「＜傘の色 色=“黄”＞」を付与するため、表記揺れの吸収や集計、解析を容易にする効果がある。2019 年１月時点で 40 点の日本語図鑑および論文から 1,528 分類群、89,361 件の CCC を収載済みである。

本稿執筆時点で本プロジェクトは未だ試行段階にあるが、これまでの作業で、データの作成に要する時間と労力が過大であることに加え、図鑑や論文の内容をそのまま切り出すだけとはいえ、作業者ごとの成果物のばらつきが予想以上に大きいことも明らかになった。そこで、光学的文字認識（OCR）により図鑑のスキャン画像を文字列データに変換し、その後辞書照合などの処理を行うことによって、ボタンを押すだけで作業を行うことができる「入力支援ツール」を開発した（図２）。これを用いれば、手作業による文字入力が最小限に抑えられ、ある程度熟練すれば１種あたり２分未満でデータを仕上げることが可能になる。

このデータベースを用いたインタラクティブ検索表も開発中であり、既に「傘の色」、「襞の密度」、「柄の表面性状」等、様々な形質に基づく検索が可能になっているほか（図３）、画像認識による種判定との連携の構想もある（図４）。

本稿では詳細な議論は避けるが、一般論として、画像認識による候補種の絞り込みは実用化が十分視野に入っており、2018 年にスマートフォンアプリ「LINNÉ LENS （https://lens.linne.ai/ja/)」が水族館および動物園の生物の高精度認識を可能にしたことを踏まえても、今後画像認識は「当たり前の技術」になっていくと思われる。ただし、画像認識による種同定に関しては、いわゆる人工知能全般にあてはまるとおり、判定結果の精度だけでなく、むしろ説明可能性（explainability）が重要だと筆者は考えている。言い換えれば、ユーザー（人間）が事前の知識や経験の多寡に拠らずに正しい同定結果に到達するためには、単なる候補種一覧のアウトプットのみならず、形質比較表や写真などの二次的な資源が必須である。「ポスト機械学習・画像認識」の段階では、形質データの充実度が同定の精度を左右するともいえるだろう。

TraitBank に日本語文献のデータが収載されることは当面期待できないことから、それとは独立した取り組みとしてデータの蓄積を続けていくことは、将来に資する事業になりうると考える。ただし、本プロジェクトの成果物をオープンデータとして公開するには著作権上の問題もあることから、将来的には適切な運用法を議論する必要性が生じると思われる。形質データをファクトデータと

傘のサイズ★4-10 cm　子実体のサイズ★4-10 cm	caprands (Scop .) Fr .「傘は4-10 cm
傘の形状★まんじゅう形からやや中高の平らとなり　子実体の形状★まんじゅう形からやや中高の平らとなり	まんじゅう形からやや中高の平らとなり
傘の色★灰褐-暗灰褐色　子実体の色★灰褐-暗灰褐色	表面は灰褐-暗灰褐色で
傘の表面性状★湿っているとき多少粘性をあらわすが乾き　傘の数量★多　傘の質感★粘性　子実体の表面性状★湿っているとき多少粘性をあらわすが乾き　子実体の数量★多　子実体の質感★粘性	湿っているとき多少粘性をあらわすが乾きやすい
肉の色★白色　肉の質感★もろい　子実体の色★白色　子実体の質感★もろい	肉は白色でもろい
襞の色★白-淡クリーム色　襞の付着様式★直生状垂生し　子実体の色★白-淡クリーム色　子実体の付着様式★直生状垂生し	襞は白-淡クリーム色で直生状垂生し
襞の密度★疎　子実体の密度★疎	疎
柄のサイズ★5-12 cm　子実体のサイズ★5-12 cm	柄は5-12 cm ×1-2 cm
柄の形状★上下同大または根もとが細まり　柄の位置★根もと　子実体の形状★上下同大または根もとが細まり　子実体の位置★根もと	上下同大または根もとが細まり
柄の表面性状★表面は粘性なく　子実体の表面性状★表面は粘性なく	表面は粘性なく繊維状

図 2.　入力支援ツール
記載文データを機械的に読み取って形質候補を表示する。ボタンを押すとクリップボードにデータがコピーされ、そのままエクセルファイルに貼り付けることができる。

図 3.　インタラクティブ検索表
左に CCC 一覧、右に種名一覧が表示されており、選択した CCC を全て持つ種のみが色付きで表示される

図 4. 画像認識による種名判定を形質データベースと組み合わせた例
右の写真（ナラタケ）の判定結果それぞれについて、他の候補種が持っていない CCC から優先的に表示している。

みなすことで著作物に当たらないと解釈可能か、あるいはフェアユースの観点から一定の利用行為が認められるか、などが論点となるだろう。

3．アノテーション支援ツール tagtog による記載文構造化データベース構築

筆者は千葉菌類談話会通信 32 号（2016 年）に投稿した記事で、文献の自動索引について、「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（中略）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」と述べたが、まさにその CRF の手法を利用した、誰でも簡単に文章にアノテーション（注釈）を付与できるサービスが登場した。それが「tagtog（https://www.tagtog.net）」である。色ペンでマーキングする感覚で、単語や文章にエンティティを付与することができ、それが即座に機械学習の結果に反映される仕組みである。これにより、これまで過大な労力を要していた英語記載文の注釈付け作業を格段に省力化することが可能になった（日本語にも対応しているというが、まだ試していない）。

筆者は tagtog を用いて、商用目的での派生作品頒布も可能な CC-BY ライセンスが付与された記載文（主に学術雑誌「MycoKeys」由来）を対象に、10 種のエンティティを用いて構造化作業を進めている。エンティティの種類は「色」「形状」「サイズ」「位置」「数量」「表面性状」「構造」「発達」「質感」「有無」であり、完全ではないものの、前述した「CCC」のカテゴリに対応している。図 5 は注釈付けを完了した段階のスクリーンショットであるが、文章をアップロードした時点で機械学習により自動的にエンティティが付与されるので、実際の作業は微修正程度である。

tagtog は WebAPI を提供しており、これを利用して特定のアノテーションが付与されたデータを自在に抽出することが可能である。図 6 および図 7 は、それぞれ「モリノカレバタケ属菌の傘直径」「カラハツタケ属菌の柄のサイズ」を可視化したものであるが、適切なプログラムを組めば、このような図はボタン一つで作成可能であり、解析対象の分類群や形質を切り替えることも容易である。tagtog は高額なサービス（下位の MEDIUM プランでも月額 250 ユーロ）なので個人での利用はあまり現実的でないが、WebAPI の使用は無料版でも特に制限がないようである。本稿執筆時点で、筆者が現在作成しているコーパスは間もなく全ユーザーに共有される予定になっているため、WebAPI 経由で自由に使えるようになるほか、ごく小さな無料枠の範囲内（10 文書/月）であれば、手元にある記載文の自動アノテーションを手軽に試すことも可能になると思われる。

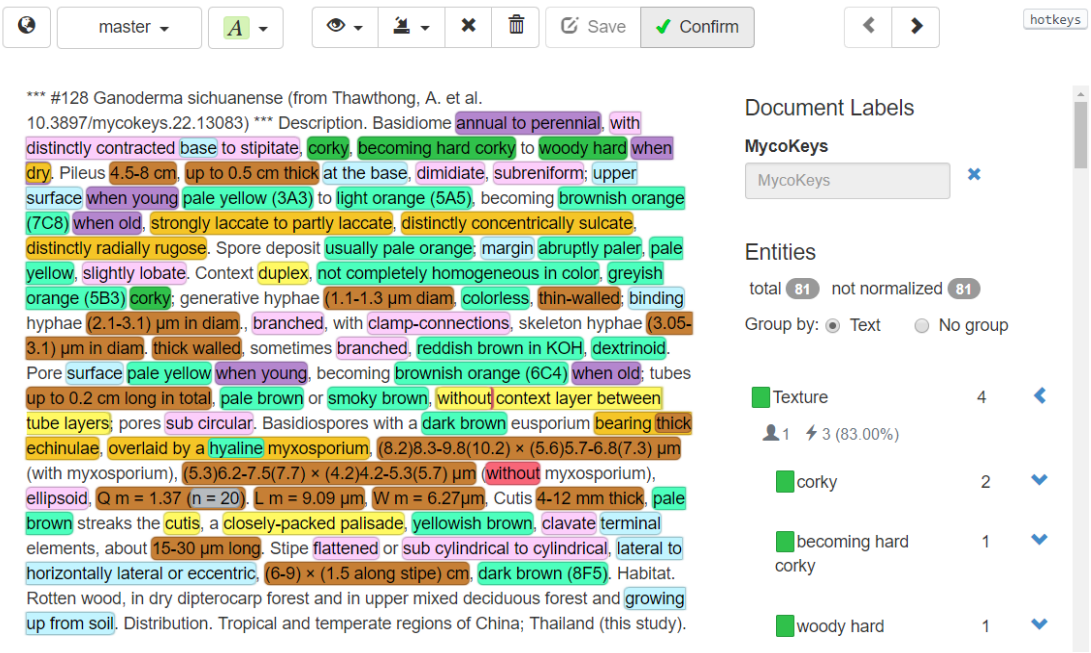

図 5. 注釈支援ツール tagtog のスクリーンショット
色や形状などを異なる色で塗り分けている。新しいデータをアップロードすると、それまで付与したデータを基に、自動で塗り分けが行われる。

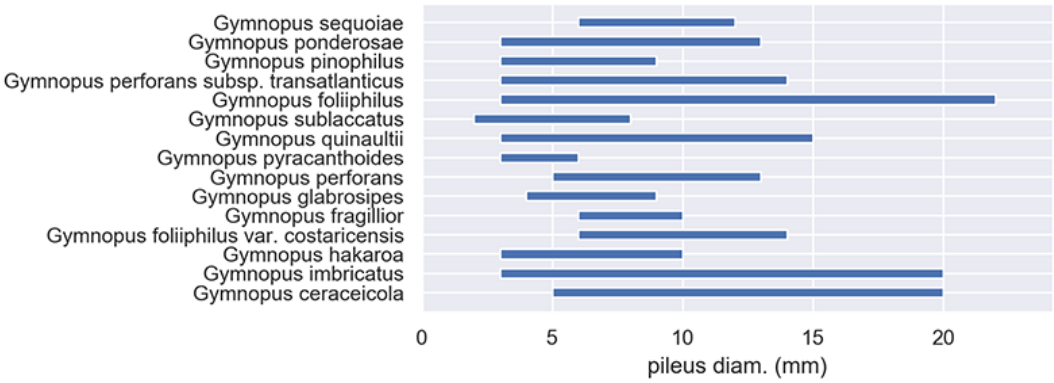

図 6. tagtog のコーパスに登録した記載文データから WebAPI 経由で「モリノカレバタケ属」の「傘直径」データを抽出して可視化

tagtog は当然ながら菌類の記載文の構造化を目的に開発されたツールではなく、当初はショウジョウバエのデータベース「FlyBase」における遺伝子等の固有表現抽出のために用いられた。実際のところ、日々発表される膨大な論文から化学物質、疾患、遺伝子、タンパク質やその相互作用などを抽出して構造化データに変換する需要は高く、「バイオキュレーター」と呼ばれる専門職が作業に携わっている例もある。2016 年に、生物医学文献の主要なオンラインデータベースの一つである「Europe PubMed Central」

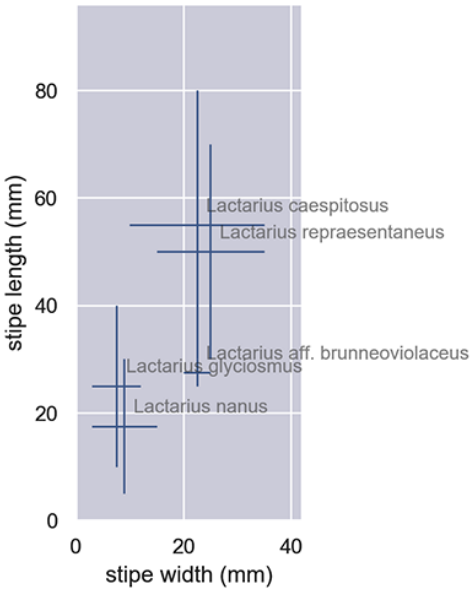

図 7. tagtog コーパスに登録した記載文データから WebAPI 経由で「カラハツタケ属」の「柄の長さ」および「柄の幅」データを抽出して可視化した図

が「SciLite」というシステムを開発し、検索性向上の目的で論文フルテキストからテキストマイニングで得た注釈情報を提供し始めたという動向を踏まえても、今後も論文からの情報抽出の重要性は高まっていくとみられる。菌類記載文の構造化に関しても、理想的には筆者のようなアマチュアではなく、バイオキュレーションの専門家が取り組むべき課題だと思うが、現段階ではこのような地道な作業は研究者の実績に繋がりにくいと思われる。今後この種の貢献が正当な評価を得るためにも、形質データベースの有用性を積極的に示していく必要があるだろう。

まとめ

以上、菌類の形質データベースに関連する３つの話題を紹介した。特筆すべきは、これらが互いに連携可能という点である。特に、これまでほとんど世界に発信されてこなかった日本語記載文の情報が構造化を経て「翻訳」されれば、データの可用性は大きく広がると思われる。複数の API を組み合わせて新たなアプリケーションを作成することをマッシュアップ（mashup）というが、きのこ（mushroom）の分野でも同様に新たな価値が創出されることが期待される。そして、そのような成果を挙げるために、良質なデータベースの存在が前提となることは論を俟たない。

形質データベースの整備は多大な時間と労力を要し、かつ学問が存続する限り決して終わることのない事業である。地道な作業の絶え間ない継続は当然ながら、それに加え、常に最新の知見に向けてアンテナを張り、それらを取り入れつつ効率化・省力化の可能性を追求する姿勢も必要であろう。

❐

― 正誤表 ―

千葉菌類談話会通信 34 号に掲載された記事『菌類オントロジー構築に向けたマイナータームーマイニング』について、以下の誤りがあったため訂正する。

● 25 ページ右段 下から 2 行目
（誤）interascal 子嚢内―〔形〕
↓
（正）interascal 子嚢間―〔形〕

● 27 ページ右段 11～12 行目
（誤）Schaeffer reaction シェッファー反応（テングタケ類の呈色反応の一種）
↓
（正）Schaeffer reaction シェッファー反応（ハラタケ類の呈色反応の一種）

● 28 ページ左段 32 行目
（誤）woodchip ウッドチップ，おが屑
↓
（正）woodchip ウッドチップ，木片