菌類オントロジー構築に向けたマイナーターム・マイニング

中島　淳志

近年、自然言語処理の手法の発展とともに、人間の目では到底読み切れない大量の文書を機械的に解析して有効活用に繋げる動きが広がっており、それに関連して様々な学問分野で「オントロジー」の構築が進んでいる。

オントロジーは特定の分野における概念およびその関係性（同義、類義、包含など）が記述されたデータベースと定義できる。例えば、人間は「鼠色」が「灰色」に近い概念であること、「ドクツルタケ」が「テングタケ属」に含まれ、それがさらに「テングタケ科」に含まれること、「子実体」が「菌糸体」から生じること、「厚膜胞子」が「厚壁胞子」の旧称であること…など、概念の関係性を容易に知りうるが、それらは全て、機械に単に文字列データを与えても「解釈」させることはできない情報である。その実現はオントロジーを利用した「セマンティック技術」によって可能になる。オントロジーに基づく注釈（アノテーション、タグ付け）を施してデータを「構造化」し、概念の関係性を明示することで、データを機械可読な形に変換することが可能になる。

構造化されたデータは、情報検索、関係性抽出、推論、可視化など、様々な場面で威力を発揮する。また、情報処理の速度および量（スループット）が向上することで新たな知見の効率的獲得にも繋がる可能性がある。オントロジーに特化した検索手段もあり、例えばSPARQL（スパークル）という問い合わせ言語を用いると、概念間の関係を「主語（S)」「述語（V)」「目的語（O)」の要素を任意に指定することで、従来の方法では困難だった柔軟かつ高度な検索が可能になる（※脚注参照)。類似概念の統合により曖昧性が軽減されることも主要な効能の一つであり、これにより異なる学問分野、異なるデータベースとの情報の相互運用が容易になることも期待される。つまり、オントロジーの構築は文書データが持つ潜在的な価値を高め、可能性や選択肢を広げることに繋がるのである。

菌類分類学の分野では「記載文」という、長年に亘り蓄積された、そして今後も学問が存続する限り蓄積されていくであろう、莫大な文章情報が存在する。しかし、それを顧みるのは当該分類群の研究者にほぼ限られ、価値を知られないまま埋もれてしまっているレガシーデータの量は未知数である。その資源を有効活用する上で、オントロジーに基づくアプローチは有望な候補になりうると考える。

オントロジーは生物学分野においては、主に遺伝子の機能を記述する目的で発展してきたが（Gene Ontology, GO)、近年は分類学分野においても、記載文を構造化して自動的に注釈を施す試みがなされている（Cui, 2010)。例えば、植物学分野では Plant Ontology（PO）、細菌学分野では MicrO、ク

モ学分野では Spider Ontology…など、分野毎にオントロジーによるマッピングを取り入れた先行研究がある（Walls et al., 2012; Blank et al., 2016; Cui et al., 2016）。

菌類の既存オントロジーとしては Fungal Gross Anatomy Ontology（FAO）が挙げられるが、用語数は僅か 100 程度である。また、細胞生物学や遺伝学の視点で編纂されており、分類学的視点における重要語はほとんど収載されていない。さらに、オントロジーの不在のみならず、それが一因かどうかは定かでないが、菌類の記載文を対象とした生物多様性情報学的研究すらも、筆者が文献を渉猟した限りでは見出すことができなかった。

オントロジー構築の基盤となるのは体系的・網羅的な語彙集である。理想的には記載文で用いられる全ての語を漏れなく照合することが望まれるが、ゼロベースの作成は現実的ではなく、既製の用語集を活用し「巨人の肩に立つ」方策が最善手と思われる。

日本菌学会編「新菌学用語集」（2014 年）は 1996 年に出版された「菌学用語集」の改訂・増補版であり、菌学全般および関連領域の用語が多数収録されている。そこで本研究では、テキストマイニングの手法により、「新菌学用語集」を語彙収集のベースラインとして捉えた際の網羅性および利用可能性を検証するとともに、「新菌学用語集」未掲載のマイナーな単語の一斉抽出を試みた。

未掲載語マイニング

筆者が様々な文献や Web サイトなどから蒐集した 17,304 件の記載文を解析に供した。まず、「新菌学用語集」の内容を人手で編集し、単数形と複数形に分割し、表記揺れを処理するなどの整形を施して 9,509 語を得た。これに Python の Natural Language Toolkit（NLTK）コーパスの「words」および「wordnet」から入手した単語集に inflect パッケージを用いて複数形を付与した語集合（一般語の網羅的集合と仮定）を加え、計 666,928 語からなる検索語集合を得た。各記載文に対して検索語集合の全ての語を照合し、照合されなかった単語から「数字を含む語」を除外した。次に、この集合に含まれる単語を計数して頻度 10 未満の語を除外した 946 語について、筆者が一つずつ精査して採否を検討した。最終的に、以下の 239 語を得た。

抽出された用語一覧

訳語は可能な限り「新菌学用語集」における類似語や信頼性の高い日本語論文などの資料を参考に正確を期したが、あくまで非専門家の一個人による仮の訳であることを留意していただきたい。

A

achlyoid　遊走子嚢が）Achlya 型―〔形〕
acropetally　求頂的に―〔副〕
acrophysalidic　頂端膨張菌糸―〔形〕
aculeus (pl. -ei)　刺
acyanophilic　コットンブルー非染色性, 非シアノフィリック―〔形〕
adelophialide　アデロフィアライド（※基部に隔壁を欠くフィアライド）
ampuliform　アンプル形―〔形〕
amyloidity　アミロイド性
anastomosed　吻合した〔形〕
annellated　（分生子形成細胞に）アネレーションがある〔形〕
apapillate　乳頭突起でない, 乳頭突起を欠く〔形〕
apedicellate　柄を欠く〔形〕
apotheciate　子器形成性―〔形〕
approx.　およそ
ascal　子嚢―〔形〕
ascomal　子嚢果―〔形〕
ascomatal　子嚢果―〔形〕
assimilated　資化―〔形〕

attaching　付着する〔形〕
attenuating
　(柄、細胞などが) 先細り―〔形〕

B

barrel(-)shaped　樽形―〔形〕
basipetally　求基的に―〔副〕
biatorine　ビアトラ型―〔形〕
bifusiform　両紡錘形―〔形〕
bilayered　二層―〔形〕
bisporic　2 胞子性―〔形〕
bisterigmate　(担子器が) 2 胞子性―〔形〕
boletoid　イグチ型―〔形〕
brachybasidium (pl. -ia)　短担子器
bruised　傷ついた〔形〕(※子実体の変色性を表現する時などによく用いられる)

C

caespituli
　(cercosporoid fungi の分生子柄の) 叢
carbonis(z)ed　炭化―〔形〕
catenescent　鎖生―〔形〕
caulobasidium (pl. -ia)
　柄の担子器 (※定訳なし？)
caulocystidial　柄シスチジア―〔形〕
caulocystidioid　柄シスチジア状―〔形〕
caulohymenium　柄子実層
cauloparacystidium (pl. -ia)
　柄類シスチジア
cephalodium (pl. -ia)　頭状体
charred　(殻壁などが) 炭化―〔形〕, (発生基質が) 焼け焦げた〔形〕
chasmothecium (pl. -a)　閉子嚢殻
cheiloleptocystidium (pl. -ia)
　縁薄壁シスチジア
cheilomacrocystidium (pl. -ia)
　縁マクロシスチジア
cheilopseudocystidium (pl. -ia)
　縁偽シスチジア
chlorococcoid
　(地衣類の共生藻が) 緑藻―〔形〕
clamped　クランプを持つ〔形〕
clampless　クランプを欠く〔形〕
cleistothecial　閉子嚢殻―〔形〕
cleistothecioid　閉子嚢殻状―〔形〕
clubshaped　棍棒形―〔形〕
coelomycetous　分生子果不完全菌―〔形〕
collapsed　崩壊した〔形〕
collapsing　崩壊する〔形〕, 崩壊性〔形〕
concolourous　同色―〔形〕
congophilous
　コンゴーレッド染色性―〔形〕
conidiogenous　分生子形成―〔形〕
conidiomatal　分生子果―〔形〕
cystidial　シスチジア―〔形〕
cystidiate　シスチジア―〔形〕
cystidioid　シスチジア状―〔形〕
Czapek-Dox Agar, Czapek Agar, CzA
　ツァペック-ドックス寒天培地

D

Dalmau plate method　Dalmau 平板法
decaying　腐朽―〔形〕
deliquescing　液化―〔形〕
dextrinoidity　デキストリノイド性
dimpled　小さな窪みのある〔形〕
disarticulating
　分節する, 離断する〔形〕
discolo(u)ring　退色する〔形〕
disintegrating　非一体―〔形〕
distoseptum (pl. -ta)　異隔壁

E

echinula (pl. -ae)　小刺, 細刺
ectomycorrhizal　外生菌根―〔形〕

ectostromatic　外子座―〔形〕
effused　平らに広がった，拡散した〔形〕
eightspored　8 胞子性―〔形〕
elongating　伸長する〔形〕
emarginated　陥入―〔形〕
encrusting　(結晶などに) 覆われる〔形〕
enteroblastically　内生出芽型―〔副〕
enveloped　包まれる〔形〕
epicuticular　表皮上層，上表皮―〔形〕
epihymenial　子実上層―〔形〕
epinecral layer　ネクラル上層
esorediate　粉芽を欠く〔形〕
euamyloid　(真の) アミロイド―〔形〕
eucapillitial
　(真の) 弾糸，細毛体―〔形〕
eucapillitium　(真の) 弾糸，細毛体
eustromatic　真の子座―〔形〕
exannulate　つばを欠く〔形〕
exuded　浸出 (滲出) した〔形〕
exuding　浸出 (滲出) する〔形〕

F

fanshaped　扇形―〔形〕
fermented　発酵した〔形〕
FeSO4, iron sulfate　硫酸鉄
flared　(傘の縁部などが) フレア状―〔形〕
flaskshaped　フラスコ形―〔形〕
fragmenting　断片化する〔形〕
fruitbody (pl. -dies)　子実体
furcated　叉状―〔形〕
fusing　融合する，癒合する〔形〕
gasterocarp　(腹菌類の) 子実体

G

GBIF　地球規模生物多様性情報機構 (※日本ノードは JBIF)
gelatinis(z)ed　ゼラチン化―〔形〕
gelatinis(z)ing　ゼラチン化―〔形〕
genbank　ジェンバンク (NCBI の塩基配列データベース)
germinated　発芽した〔形〕
germinating　発芽する〔形〕
germpore　発芽孔
gleoplerous　粘質―〔形〕(※「新菌学用語集」の gloeo-と同じ)

H

hamathecial　子嚢果内菌糸系―〔形〕
hemisphaerical　半球形―〔形〕
heterodiametrical　異径―〔形〕(※イッポンシメジ類の胞子の形状を表現する時などによく用いられる)
hollowed　中空―〔形〕
HPLC, high-performance liquid chromatography
　高速液体クロマトグラフィー
hymenophoral　子実層托―〔形〕
hypocutis　下皮層

I

I+ / J+ / I- / J-　ヨウ素陽性/陰性 (アミロイド/非アミロイド)
IKI+/IKI-
　ヨウ素ヨウ化カリウム陽性/陰性
imbedded　埋生―〔形〕
inaequilateral　左右不等―〔形〕
indel　インデル (塩基配列の挿入および欠失)
indextrinoid　非デキストリノイド
inequilaterally　左右不等―〔形〕
inflating　膨大―〔形〕
inoculated　接種した〔形〕
insititious
　(柄の基部に) 菌糸体を欠く〔形〕
inspersed　(顆粒、液滴などを) 含む，散在する〔形〕
interascal　子嚢内―〔形〕
intermingled　混在―〔形〕

interspersed　散在―〔形〕
intertwined　絡み合った〔形〕
interveined　(ひだが) 脈間―〔形〕

L

lichenized　地衣化した〔形〕
Lugol's solution　ルゴール液

M

macrobasidium (pl. -a)　大担子器
microcharacter　肉眼的形質
micromorphology　肉眼的形態
macroscopy　肉眼的観察
macula (pl. -ae)　斑点
manured　(生息環境が) 施肥―〔形〕
Melzer　メルツァー試薬
metuloidal
　メチュロイド, 冠石灰嚢体―〔形〕
microbasidium (pl. -a)　小担子器
micromorphology　顕微鏡的形態
MLZ+ / MLZ-　メルツァー液陽性/陰性
monophialide　モノフィアライド (※ポリフィアライドに対して胞子が生じる部位が単一のフィアライド)
multilayered　多層―〔形〕
multispored　多胞子性―〔形〕
mummified　(昆虫病原菌の宿主などが) ミイラ状―〔形〕

N

NaCl, sodium chloride　塩化カルシウム
NaOH, sodium hydroxide　水酸化ナトリウム
nested　入れ子状―〔形〕
nonamyloid　非アミロイド (※新菌学用語集に"non-amyloid"および"inamyloid"有)
non(-)dextrinoid　非デキストリノイド
non(-)gelatinis(z)ed
　非ゼラチン化―〔形〕
non(-)marginate　縁取りを欠く〔形〕
non(-)ostiolate　孔口を欠く〔形〕
non(-)papillate　乳頭突起でない, 乳頭突起を欠く〔形〕
non(-)striate　条線を欠く〔形〕
non(-)stromatic　非子座―〔形〕

O

obovoidal　倒卵状―〔形〕
odontioid　コメバタケ型―〔形〕
oleiferic　油管―〔形〕
omphalinoid
　ヒダサカズキタケ型―〔形〕
overhanging　張り出す―〔形〕

P

paracapillitial　細毛様体―〔形〕
paraplectenchymatous
　異形菌糸組織―〔形〕
pedicillate　有柄―〔形〕
penetrated　侵入した〔形〕
percurrently　(分生子形成細胞が) 貫生伸長―〔副〕
peridiopellis　殻皮表皮
pers. comm., personal communication
　私信 (※未公表の伝聞情報であることを指す)
phenoloxidase　フェノールオキシダーゼ, フェノール酸化酵素
phloxine　フロキシン
pigmented　有色の, 色付いた〔形〕
pileal　傘―〔形〕
pileitrama　傘実質 (※新菌学用語集に"pileus trama"有)
placodioid　鱗殻状―〔形〕
pleated　(傘縁部などが) 扇畳み, プリ

一ツ状〔形〕
plectenchymatous　菌糸組織―〔形〕
pleuroleptocystidium (pl. -ia)
側薄壁シスチジア
pleuromacrocystidium (pl. -ia)
側マクロシスチジア
pleuropseudocystidium (pl. -ia)
側偽シスチジア
probasidial　前担子器―〔形〕
proliferated　(分生子形成細胞などが)伸長する〔形〕
prosoplectenchymatous
繊維菌糸組織―〔形〕
pruina　粉霜, 微粉
pseudocollarium　偽襟帯
pseudodimitic　偽二菌糸型―〔形〕
pseudoparenchymatic　偽柔組織―〔形〕
pseudorhizal　偽根―〔形〕
pseudorrhiza
偽根 (※pseudorhiza の別綴り)
pseudostipitate　偽柄―〔形〕
pseudothecial　偽子嚢殻―〔形〕
pycnidioid　分生子殻―〔形〕

Q

quaterverticillate　四輪生―〔形〕(※アオカビ類などの分生子柄の表現に用いられる)

R

radish　ダイコンのような〔形〕(※子実体の臭いなどが)
Rameales structure　ラメアレス構造 (※傘表皮の構造の一種、シロホウライタケ属の節の名称に由来。Ramealis structure[ラメアリス構造]との表記も見受けられる)
ramifying　分枝する, 分岐する〔形〕
rehydrated　吸水―〔形〕
rehydrating　吸水する〔形〕
released　(遊走子, 子嚢胞子などが)放出した〔形〕
rhizinate　偽根―〔形〕
rupturing　破裂する〔形〕

S

saprotrophic　腐生―〔形〕
Schaeffer reaction　シェッファー反応 (テングタケ類の呈色反応の一種)
sclerocystidium (pl. -ia)
スクレロシスチジア
scleroplectenchymatous
厚壁菌糸組織―〔形〕
semiimmersed　半埋生―〔形〕
sensu (人名)　(人名)が言及(定義)するところの～
Siccus-type　Siccus 型, ハリガネオチバタケ型―〔形〕(※箒状細胞からなる傘表皮)
skeletobinding　骨格結合―〔形〕
spheropedunculate　有柄球形―〔形〕
sporangiate　胞子嚢―〔形〕
sporebearing　胞子を生じる, 胞子を形成する, 胞子を含む〔形〕
sporeprint, spore deposit　胞子紋 (※新菌学用語集に”spore print”有)
sporulating　胞子形成―〔形〕
stereomicroscope　実体顕微鏡
sterigmal　小柄―〔形〕
stipititrama　柄実質
striatulate　条線を有する〔形〕
stromatal　子座―〔形〕
subtending hypha (pl. -ae)　(グロムス類の)サブテンディング菌糸 (※定訳なし?)
sulfovanilin　スルフォバニリン
suprabasal

基部の上に位置する，基底上の〔形〕
supramedian　（胞子の隔壁などが）真ん中より上―〔形〕
synasexual　シンアセクシャル―〔形〕（※複数の無性世代を持つ菌の形容に用いる、定訳なし？）

T

terminating
末端生―〔形〕，末端が～になる〔形〕
terpenoid　テルペノイド
tetrasterigmate
（担子器が）4 胞子性―〔形〕
tetrasterigmatic
（担子器が）4 胞子性―〔形〕
thick(-)walled　厚壁―〔形〕
thin(-)walled　薄壁―〔形〕
thromboplerous　凝血状―〔形〕（凝血状菌糸の）
transvenose　（ひだが）脈間―〔形〕
trebouxioid　（地衣類の共生藻が）トレボウキシア型―〔形〕

U・V・W

uncarbonized　非炭化―〔形〕
volval　つぼ―〔形〕
woodchip　ウッドチップ，おが屑

※「sub-」は「亜～」「類～」を表す頻出の接頭語であり、「subfusoid」「subcapitate」「sublageniform」など 68 語がヒットしたが、この接頭語は任意の形容詞に柔軟に付与可能であることから、上の一覧からは除外した。

※「atranorin」「salazinic acid」「usnic acid」のような地衣成分、あるいは試薬や培地などに用いられる化学物質の名称も「新菌学用語集」には収録されていないものが多いが、上の一覧からは多くを除外した。

考察

今回のマイニングで得られた用語の中には、「必須」といえるレベルの重要性を持つ菌学用語は見出されなかった。このことから、「新菌学用語集」が菌類の記載文で用いられる専門用語を十分広範に網羅していることが見て取れる。

上の一覧に挙げたのは、既に掲載されている用語の形容詞形や、記載文でよく用いられる一般語が主であるため、「重箱の隅を突いたような一覧」との印象を持つ読者もいるかもしれない。しかし、「人間にとっての実用性」と「機械可読性」は全く別の概念であり、価値観も根本的に異なっている。オントロジー設計という視点に立ち、後者を重視する筆者の立場からは、ごく微妙なスペル、ニュアンスの違いでも可能な限り全ての語を網羅し、将来の用語整理に必要な資源を可能な限り集めることが重要だと考える。

今回の解析手法は単独では決して完全ではなく、単語ベースの収集であるため熟語表現が取得できないほか、必要な専門用語が偶然に一般語と同じ綴りで除外されてしまった可能性、あるいは逆に除外された一般語が特殊な用例で専門性を帯びていた可能性も考えられる。オントロジー構築にあたっては、実作業の過程で検証を重ね、継続的に取捨選択を行う必要があるだろう。

なお、本解析では個人的に文献や Web サイトから蒐集した記載文を用いたが、本来は再現性の高い公正な解析を行うには公的データベースのデータを用いるのが望ましい。しかし、例えば MycoBank に収載されている記載文は複数の言語が混在しており、用語や概念が異なる古い時代の記載文も多く、ノイズデータも多く含まれることなどから、解析は困難であった。地衣類に関しては Botanische Staatssammlung München （“LIAS”）、Consortium of North American Lichen

Herbaria (CNALH) 等が Web 上で豊富な記載文を公開しており、ある程度の構造化がなされているため、MycoBank よりも解析が容易である。今後、菌類学におけるオントロジーの必要性が周知され、他の菌群でもこのような充実した資源が利用可能になることが望まれる。

オントロジーの構築および適用は分類学に革新的な変化をもたらす可能性があるが、少なくとも菌類については植物、昆虫など他の生物よりも大きく立ち後れているのが現状である。現段階から遺伝子や疾患のような洗練されたオントロジーと同等の質のものを作り上げることは容易ではなく、道のりは果てしなく長いが、今回の解析で得られたデータが僅かでも一助になれば幸いである。

脚注

※一例として「<イグチ (S)>が<トビムシ (O)>に<食べられる (V)>」という関係を挙げると、O と V を固定して S を検索することで「トビムシに食べられる菌の一覧」、S と V を固定して O を検索することで「イグチを食べる昆虫や動物の一覧」、S と O を固定して V を検索することで「イグチとトビムシの関係の一覧」を検索することができる。

引用文献

Blank, CE. et al., 2016. MicrO: an ontology of phenotypic and metabolic characters, assays, and culture media found in pro-karyotic taxonomic descriptions. Journal of Biomedical Semantics. DOI:10.1186/s13326-016-0060-6

Cui, H. 2010. Semantic annotation of morphological descriptions: an overall strategy. BMC Bioinfomatics. DOI:10.1186/1471-2105-11-278

Cui, H. et al. 2016. Introducing Explorer of Taxon Concepts with a case study on spider measurement matrix building. BMC Bioinformatics. DOI:10.1186/s12859-016-1352-7

Walls, R. et al. 2012. Mapping of glossary terms from the Flora of North America to the Plant Ontology enhances both resources. ICBO 2012. http://ceur-ws.org/Vol-897/poster_6.pdf

❒

［図書紹介］

■しっかり見わけ観察を楽しむ きのこ図鑑

著者：中島淳志、写真：大作晃一、監修：吹春俊光、
発行所：ナツメ社、2017 年 10 月 5 日発行、小 B6 判、
320 頁、ISBN: 978-4-8163-6303-0、1300 円＋税

見た目から探せる直感インデックス、原寸大の識別用写真、発生場所・毒の有無のアイコン表示など、見分けやすさ・探しやすさにこだわり、309 種を掲載したハンディタイプ図鑑。さらに約 1 万 5 千件の学術データを収集し、医学分野で使われる解析手法を応用、グループごとの大きさ、色、発生時期の傾向がビジュアルでわかる。マニアックな解説も見どころ！ ㈫