生物学・生命科学分野におけるオンライン文献検索

第2回 BIOSIS (2) ディスクリプタ[†]

山 中 啓
(筑波大学応用生物化学系)

6. 同義語・類義語

化学反応にも同義語が多いが、生物学分野にも同義語が多く、同義語の集合をまず作ることを念頭におくことが望ましい。二、三の例をあげて説明する。

(a) 分解, 代謝 生物の示す基礎的反応の一つであるために、同義語が非常に多い。たとえば、

degradation, biodegradation, decomposition, dissimilation, metabolism, catabolism, consumption, breakdown.

これらの用語は1文献に2種以上含まれることは、ほとんどなく、主として報文のタイトルからキーワードがとられているので、報文の著者が用いた用語のみがキーワードに採用されているようである。したがって、まずこれらの集合の和を作り、次のキーワードと論理積にするのが望ましい。

(検索例 1)

Degradation	1997 件
Decomposition	586
Dissimilation	6
Metabolism	9437
Breakdown	341
Degradation+Decomposition	2567 (570 増)
+Dissimilation	2573 (6 増)
+Metabolism	11913 (9440 増)
+Breakdown	12235 (295 増)

検索例1でわかるように、これらの用語は検索上の相

互関連がなく、単純和と論理和がほとんど同じである。これはこのデータベースが自然語で検索できるようになっているためであって、シソーラスで用語管理(統制)がされていないからである。もちろん、システムに辞書機能があって、“DEGRADATION”と入力すると、自動的に上記の関連語についての検索を実行してくれれば問題はかなり解消するであろう。これらの用語の省略形(DEGRADATIONはDEGRDN)はCASにあり、これも含める必要があるが、BIOSISにはない。

この例は多いので、検索もれを出さないため、とくに留意すること。例: uptake, incorporation, assimilation, fixation.

(b) レンニンとレンネット 凝乳酵素の正式名はレンニンであるが、産業ではレンネットも使用されているので、両者の和を作る。

(検索例 2)

Rennin	20 件
Rennin@	22
Rennet	54
Rennet@	59
Rennin@+Rennet@	81

したがって、rennin と rennet の2つのキーワードが同一文献中に含まれないので、和の集合が必要である。(この集合を用いた検索例を14に示した。)

シソーラスで管理されているデータベースでは、この混乱はない。

7. 複 数 形

検索上は単数形と語尾にsのついた複数形は別々の用語である。とくにラテン形で語尾の変るものは注意する。

(検索例 3)

[†] A Practical Guide for Information Retrieval on Biology and Life Science. Part 2. BIOSIS (2) Descriptor
Kei YAMANAKA (Institute of Applied Biochemistry, The University of Tsukuba, Sakura-mura, Ibaraki, 305)

(a) 微生物

Microorganism	560 件
Microorganisms	944

この2つの文献集合を作るためには、前方一致をとる。

Microorganism@	1498
----------------	------

この場合、この2語以外に用語はなく、前方一致で完全な論理和の集合が作れる。しかし、微生物には、ほかに microbe という用語があり、これとの関係も調査しておいたほうがよい。前記の同義語、類義語はおのこの単数形、複数形の和をとるのが望ましい。

Microbe	382 件
Microbes	94
Microbe+Microbes	467
Microorganism+Microorganisms	1498
Microorganism+Microorganisms +Microbe+Microbes	1965

これにさらに microbial を加えたほうが集合としては、より完全なものになるであろう。

(b) 抗体

Antibody	5951 件
Antibodies	2889
Antibody+Antibodies	8585

(c) ミトコンドリア

Mitochondria	2759 件
Mitochondrion	47
Mitochondri@	4586
Mitochondria+Mitochondrion	2806

(d) 線虫

Nematode	753 件
Nematodes	568
Nematoda	511
Nematode+Nematodes+Nematoda	1787
Nematod@	1819

Nematoda は線虫綱であるが、これを含めて、nematod@ の前方一致をとると十分で、ノイズは少ない。BC 51300 との関係については、次回に説明する。

(e) かび

Fungus	805 件
Fungi	2080
Fungal	1036
Fung@	5039

Mold	346
Fung@+Mold	5346

Fungus と fungi のほか、形容詞形を見逃すことはできない。不要な語の混入を覚悟で fung@ とするのも一つの方法で、得られた集合のなかには fungicidal, fungicide 等も含まれるが、検索のための集合としてはこのほうがよい場合もある。さらにかびには別に mold (英語のデータベースでは mould) もある (ただし件数は多くない)。

(f) 菌糸

Myceli のブラウジングの出力結果 (一部)

Mycelia	113 件
Mycelial	160
.....	
Mycelium	184

Mycelia@ の前方一致にすると、無関係の語の入ってくるのが防げるが、mycelium を別に加える必要がある。

Mycelia@	274 件
Mycelia@+Mycelium	455
Myceli@	463

Myceli@ の集合には、mycelioid, mycelis, mycelites 等も含まれるが、件数が少ないのであまり問題にならない。

8. 前方一致

前方一致検索は非常に有用であるが、どの範囲の前方一致にすれば最もよいかは、前もってブラウジングをして決めるのも一法である。

(検索例 4)

Phospho@	15012 件
Phospha@	15538
Phosphat@	14071

9. 複合語

(a) 赤潮 赤潮を red tide と water bloom の2つで検索した。

(検索例 5)

Red	4608 件
Red-tide	0
Redtide	0
Water	11952
Water-bloom	0
Waterbloom	0

となったので、red * tide あるいは water * bloom で

検索しなければいけないことがわかる。

Red * Tide 34 件

Water * Bloom 24

この検索結果を知るため、以下の検索をし、その結果を出力した。

Red * Tide * Plankton@ 8 件

検索された8件の文献のうち7件は red tide をキーワードにもち、内容は赤潮（エクアドル2件、フランス1件、日本2件、湧昇流1件、ペルー1件）で、残りの1件は red sea の climate tide であった。該当文献の D（キーワード）をみると、すべてに red tide が含まれているが、water bloom は全くなく、そのほかに赤潮を示すキーワードもなく、この検索法は正しい。

Water * Bloom * Plankton@ 10 件

の出力内容はすべて水の華であったが、D（キーワード）に含まれる用語はまちまちで、red tide は全くない。plankton bloom 4件で後は各1件であった。algae bloom, algal bloom, autumn bloom, water bloom, grazing bloom.

(b) 大豆 大豆は soybean か soy bean かについて調べた。

(検索例 6)

Soy 500 件

Bean@ 1523

Soy * Beans 12

Soybean 2357

Soybeans 356

したがって、大豆の場合は、(soybean+soybeans) か、または soybean@ とすれば適正な集合が得られる。

(c) 誘引性 昆虫の性誘引性の検索式を検討した。

(検索例 7)

1. Attractant 151 件

2. Sex-attractant 0

3. 1 * Sex 76

4. 1 * Sex@ 111

5. 4 * Pheromone@ 42

6. 4 * Insect@ 5

7. 4 * Cockroache@ 2

6, 7を出力すると、いずれも適合文献であったが、該当件数がもう少し多くなるような検索式のほうが安全で

ある。

(d) 大豆トリプシン・インヒビター 大豆は1語であるので（検索例 6）、これは3語に分割される。

(検索例 8)

1. Inhibitor 2692 件

2. Inhibitor@ 8050(注)

3. Inhibitors 1995

4. 1+3 4587

5. 4 * Trypsin 610

6. 5 * Soybean 204

検索用語の順番は、言葉の語順に従う必要は全くない。soybean, trypsin には複数その他の語が考えられないので、inhibitor のみ複数形を含めるように、先に集合を作り、その集合に論理積をかけてゆくのが最もよい。実際の検索式は一度でできる。

(Inhibitor+Inhibitors) * Trypsin * Soybean

(e) 窒素固定 これは、nitrogen * fixation でも fixation * nitrogen でも意味が全く変わらないので、該当件数は多くなる。検索の目標を定めて、適当な件数に正確にしぼり込む必要がある。

(検索例 9)

1. Fixation 1880 件

2. Fixation * Nitrogen 591

さらにしぼりこむと、たとえば、

3. Fixation * Nitrogen * Nitrogenase 119

4. 3 * Soil 9

出力した結果、すべてが該当文献であった。

同時に nitrogen についてブラウジングした結果、以下のような出力結果を得た。

(検索例 10) nitrogen のブラウジングの出力(一部)

Nitrogen 6815 件

Nitrogen-free 41

Nitrogen-nitrogen 2

Nitrogen-poor 1

Nitrogen-13 34

Nitrogen-14 12

Nitrogen-15 214

Nitrogenase 362

Nitrogenases 2

(注) Inhibitor のブラウジングの結果、inhibitor-drug, 2327 件, inhibitory 1557 件の2つのみが該当文献数が多かった。

Nitrogenated 3

Nitrogenous 128

この結果より、検索目的に合った項目の文献集合を作ることができる。たとえば、

Nitrogen+Nitrogen-13+Nitrogen-14+Nitrogen-15

Nitrogen@ (Nitrogen の前方一致全部の和)

Nitrogen@-Nitrogena@ (上の和集合より Nitrogenase, Nitrogenases 等の集合を除いた集合)

Nitrogen-15 (特定の一項目のみの集合)

任意の文献集合を自由に瞬時に作ることができる。

10. 複合語の分ち書き

上に述べた複合語は、本来2語以上の複合(組合せ)により意味を正確に限定したもので、文献集合を作る作業は異和感なくスムーズに行うことができる。そのほかに、われわれが通常1語として使用している用語が、機械検索を容易にするために、検索上意味のある部分に分割されている例も多い。その例を表4に示した。この表では、検索式に従って出力される該当件数の変りかたは表示していないが、左側の分割しない用語をディスクリプタとして検索すると、該当件数はいずれも「ゼロ」である。以上は一例であるが、同じような接頭語をもっている、分割されていない用語もある。たとえば、

表4 分割して検索する用語例

用 語	検 索 式
Acetylcholine	Acetyl * Choline
Adrenocortical hormone	Adreno * Cortical * Hormone
Aminopeptidase	Amino * Peptidase@
Antiserum	Anti * Serum
Antitumor	Anti * Tumor
Bacteriophage	Bacterio * Phage@
β -Glucan	Beta * Glucan@
Biodegradation	Bio * Degradation
Biotransformation	Bio * Transformation
Carbon dioxide	Carbon * Di * Oxide
Carboxypeptidase	Carboxy * Peptidase@
Corticosterone	Cortico * Sterone
Dethiobiotin	Dethio * Biotin
Dihydrochalcone	Di * Hydro * Chalcone
Dipeptidase	Di * Peptidase@
Endonuclease	Endo * Nuclease@
Endopeptidase	Endo * Peptidase@
Endotoxin	Endo * Toxin@
Enterotoxin	Entero * Toxin@
Exopeptidase	Exo * Peptidase@
γ -Aminobutyric acid	Gamma * Amino * Butyric-acid
Hemicellulase	Hemi * Cellulase@
Hemicellulose	Hemi * Cellulose
Hemoglobin	Hemo * Globin@
Hydrocarbon	Hydro * Carbon@
Isocitrate lyase	Iso * (Citrate + Citric) * Liase
Lipopolysaccharide	Lipo * Poly * Saccharide@
Monoxygenase	Mono * Oxygenase@
Neohesperidine	Neo * Hesperidine@
Peptidoglycan	Peptido * Glycan@
Phylloquinone	Phyllo * Quinone@
Phytohemagglutinin	Phyto * Hem * Agglutinin@
Polyethylene glycol	Poly * Ethylene * Glycol@
Polysaccharide	Poly * Saccharide@
Tetracyclic diterpene	Tetra * Cyclic * Di * Terpene
Ubiquinone	Ubi * Quinone@

表 5 一語として検索する用語例

(A)	Antibiotic	(B)	Antineoplastic-drug
	Biomass		Antiviral-drug
	Electrochemical		Blue-green algae
	Endoplasmic reticulum		Cell-free
	Galactosamine		Guinea-pig
	Glucosamine		In-vitro
	Hydrochemical		In-vivo
	Macrophage		
	Microorganism		Citric-acid
	Photodynamic		Pimelic-acid
	Photosynthetic		Sialic-acid
	Physicochemical		Uronic-acid

anti *biotics の集合を作ると、該当件数はゼロになるので、antibiotic @ で検索しなければいけない。このような例を表 5 に示した。またハイフンで 2 語を続けて 1 語として検索できる用語の例を (B) に示した。酸は BIOSIS では 1 語とみなして検索する。

11. 略 記 号

われわれは汎用する物質名、実験手法等の多くを略記号で表示している。しかもその略し方は国際的に統一されているものが多い。しかし、それらの略記号のすべてが検索のためのディスクリプタとして採用されているのではない。塩基類は全部 (たとえば, ATP, ADP, AMP) ディスクリプタとして利用できるが, UDPG は使用できない。糖およびアミノ酸の 3 文字よりなる略号はすべて使用できないので全綴りを用いる必要がある。したがって、ヒスチジンは histidine であって、his で検索すると、his は代名詞であることがわかる (his works, his birthday, …)。同様に man はマンノースではなく、ヒトである。塩基の A, T, G, C 等も使えない。A は冠詞であり、アデニン、アラニン、アラビノースを意味しない。T を検索すると、chloramine T, T phage T particle, T antigen, virus T 等が検索され、チミンも tRNA も検索されない。RNA, DNA は検索できるが、tRNA, mRNA は、それぞれ RNA *transfer, RNA *messenger の論理積で検索する。NAD, NADH, NADP, NADPH, FMN, FAD 等は検索できるが、CoA, PALP はだめである。実験手法として、NMR は使用できるが、ガスクロ、液クロは略記号を使用できない。

以上は一例にすぎないので、ディスクリプタとして入力したときに、該当件数がゼロまたは異常に少ないときに略記号の使用の適否を判断するとよい。

12. 応 用 例

(a) 海洋(水)プランクトンの毒性 まず toxic についてのブラウジングの結果を示す。これは toxi についてブラウジングした結果と同じであった。それは toxia, toxib という用語がないためである。

(検索例 11)

Toxic	2498 件
Toxica	1
Toxical	1
Toxicant	27
Toxicants	46
Toxicating	4
Toxication	59
Toxicities	45
Toxicity	6478
.....	...
Toxicologic	35
Toxicological	104
Toxicologically	2
Toxicologists	1
Toxicology	258
Toxi@	14414
Toxic@	9747
Toxica@	138
Toxici@	6519

と打出されたので、前方一致検索が適している。

毒性物質には一般的な用語として poison があり、これをブラウジングで打出す。

(検索例 12) poison のブラウジング

Poison	122 件
Poison-arrow	2

Poison-ivy	9
Poison-oak	1
Poisoned	67
Poisoning	706
Poisonings	38
.....	...
Poisons	32

となったので、この集合は poison@ と前方一致にするのがよい。

しかし、同一文献に toxic- と poison- がキーワードとして含まれることは少ないことは、次の検索結果より明らかになった。したがって、両者の和の集合を作る。

(検索例 13)

Toxic	2498 件
Toxic+Toxica@	2634
Toxic+Toxica@+Toxici@	9017
Toxic+Toxica@+Toxici@+Toxicol@	9453
(Toxic+Toxica@+Toxici@+Toxicol@)	
* Poison@	81
Toxic@+Poison@	10651
(Toxic@+Poison@)* Plankton@	29
(Toxic@+Poison@)* Plankton@* Phyto	16
(Toxic@+Poison@)* Plankton@* Phyto	
* Marine	3

ここで検索された3件のキーワードには poison が含まれていなかった。したがって、poison@ を省略しても全く同一検索結果が得られた。

Plankton@	2189
Plankton@* Phyto	1070
Plankton@* Phyto* Marine	97
Plankton@* Phyto* Marine* Toxic@	3

検索された3件は、重油あるいは油拡散剤 (oil-spill emulsifier, oil dispersant) による海洋プランクトンへの毒性に関するものであり、プランクトンの生産する毒性物質に関する文献が全く含まれていない。それは検索式が間違っていたためである。対策を列記する。

- プランクトンを特定の種に限定する。
- 特定のプランクトンの BC 番号の論理和を作る。
- 毒性物質を特定する。
- 毒作用を特定する。
- 被害を受ける生物種をキーワードに加える。

(b) 微生物の生産する凝乳酵素 すでにレンニンの文献集合については 6(b) の検索例2で示したので、次の検索を行った。

(検索例 14)

1. Rennet@+Rennin@	81 件
2. 1* Microbial	1
3. 1* Mucor@	11

微生物の生産するレンニンは microbial rennin と呼ぶが、microbial は検索上範囲の狭い概念であり、かえって mucor@ のほうが該当件数が多くなる。この11件を出力させたところ、すべて適合文献であったが、この検索式は適正であるか否かを調べるために、11件のディスクリプタをみた。その結果、11件中の9件に milk clotting enzymes が含まれていた。そこでこの言葉を用いて検索式を作製し直した。

(検索例 15)

1. Enzyme@	12801 件
2. 1* Milk	79
3. 2* Clotting	21

この結果と検索例 14 との論理積あるいは論理和を求めて、それぞれのディスクリプタの相互関係を調べた。

4. 3* Rennin@	13
5. 3* Rennet@	12
6. 3+Rennin@+Rennet@	89

したがって、この場合には凝乳酵素で文献集合を作るよりも、rennin と rennet の前方一致の和集合を作れば、ほとんどの確に近い文献集合を得ることができる。このなかよりかび、細菌または特定の微生物 (*Mucor*, *Aspergillus*, *Bacillus*, *Pseudomonas*) との集合を作ればよい。

(c) 排水の土壌還元

(検索例 16)

1. Land+Lands	671 件
2. 1* Disposal	24
3. 2* Waste@	9

本集合を water の積を求めなくとも、すでにその内容は処理排水の陸地還元が主であった。同様な内容を処理水の再利用を中心に考えると、次の検索ができる。

(検索例 17)

1. Renovation	15 件
2. Renovation+Reuse	53
3. 2* Water@	27

4. 3 * Waste@ 12

この検索式の組立てを反対から実施しても件数は変わらない。

5. Waste@	1063 件
6. Water@	13376
7. Water@ * Waste@	550
8. 7 * (Reuse+Renovation)	12

すなわち、該当文献集の多い集合から出発した場合と少ない集合から出発した場合の間に違いはない。また用いたキーワードが異なっても、検索例 16 と 17 には同一の文献も検索されている。

13. サーチ・ガイド (search guide)

BIOSIS では、検索のためのガイドブックとして、サーチ・ガイドを編集している。このなかのマスター・インデックス (master index) では使用頻度の高い用語あるいは概念約 12,500 語について、検索上の説明をしている。したがって、検索をより正確に行うためには、まずマスター・インデックスを調べるべきである。しかし、検索式の組立てに不馴れなときに、このインデックスで探すのは苦痛であり、オンライン検索への興味を激減させるかもしれない。

以上の点を考慮して、本編では、研究者がすぐ頭に思い浮かべるテーマより派生するキーワードを用いて、いかに検索するかの実例を示す方法をとった。そしてわれわれの使用する自然語 (free term) が検索においては不完全な用語である点を理解していただくよう努力した。本編でとりあげた用語の問題の多くは、マスター・インデックスを調べると、明記してあるものが多い。以下にその例を示す。

(a) **Biodegradation** マスター・インデックスで biodegradation の項をみる。

BIODEGRADATION

KW : BIO & DEGRADATION (400)

CC : BIODEGRADATION AND BIODETERIORATION CC 39006 (1960)

2つのキーワードによる検索か、または CC (概念コード) 39006 で集合を作る。カッコ内は頻度を示す。

(b) **複数形** インデックスに載っている用語では、KW は前方一致で検索するよう指示がされている。

(例) **RENNIN** KW : RENIN\$

\$ はトランケーション (truncation)

(c) **複合語、略記号** 複合語の分ち書き、または

一語として検索する用語は示されている。略記号もインデックスに載っている用語は使用できる。

(例) **DEOXYRIBONUCLEIC ACID**

See DNA

(d) **同義語、類義語** 毒性物質を例にして説明する。

(例) **TOXIC**

KW : TOXIC (6720) TOXICITY (22090) TOXIF\$ (620) TOXIGEN\$ (340)

See also INTOXICATION ; POISON ; TOXIN TOXICATION

KW : TOXICAT\$ (210)

TOXICOLOGY

.....

TOXIN

KW : TOXIN\$ (13030)

may appear with AFLA or ENDO or GLIO or PODOPHYLLO or TETRODO

See also POISON ; TOXIC ; TOXOID

したがって、TOXIC, TOXICITY, ... 等の集合を作るか、TOXIC の前方一致に TOXIN を加え、それに POISON を加えた集合をすぐ作ることができる。また TOXIN の項より、ENDOTOXIN の項をみると、

ENDOTOXIN

KW : ENDO & TOXIN\$ (2810)

とあり、直ちに2語の集合を作ることができる。

14. 第2回のまとめ

文献集合を作るということは表題および内容 (要旨) に含まれる用語 (キーワード) を持つ文献の集合を作ることである。その目的に使用する用語をディスクリプタと呼ぶ。ディスクリプタには自然語をそのまま用いるが、自然語のあいまいさ、不完全さを補い、より完全な検索をするためには、まずサーチ・ガイドのマスター・インデックスを参照にして、目的に最も適合したキーワードを求める。もしマスター・インデックスに目的の用語を見出すことができなかった場合には、同義語、類義語について探すこと。ただし、サーチ・ガイドを常備しているところは少ないと思われるので、まず自分のキーワードで端末操作をすることをお奨めする。もし該当件数が少ない場合は、前方一致検索をするか、ブラウジングコマンドで関連キーワードの件数を調べて検索式を組み立てる。

単数形と複数形，複合語の分割と連結に注意すれば，誌，発表年等の他の調査項目を用いた検索も可能である。
検索は十分成果をあげることができる。本論文では，デ
ィスクリプタを用いた検索式のみを示したが，著者，雑