National Bioscience Database Center

国立研究開発法人 科学技術振興機構 Japan Science and Technology Agency

生命科学系研究データ循環基盤としてのアーカイブと横断検索

○大波純一[1]、八塚茂[1]、杉崎太一朗[2]、友田史緒里[2]、牧口大旭[2]、加藤健弘[3]、井上圭介[4]、大久保克彦[3]、川本祥子[5,6]、畠中秀樹[1]

1.国立研究開発法人科学技術振興機構バイオサイエンスデータベースセンター（NBDC）、2.三井情報株式会社、3.株式会社日立製作所、4.株式会社日立公共システム、5.大学共同利用法人情報・システム研究機構ライフサイエンス統合データベースセンター（DBCLS）、6.情報・システム研究機構 国立遺伝学研究所

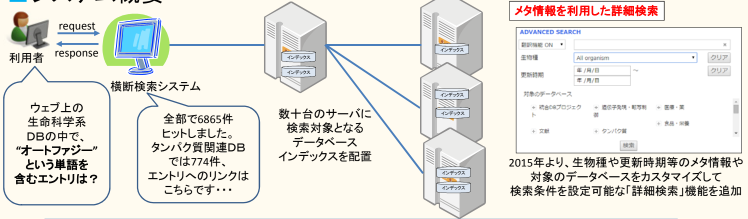
要旨

近年のオープンサイエンス推進の潮流や情報通信技術の発展に伴い、生命科学系研究データの規模は増加の一途を辿っている。バイオサイエンスデータベースセンター（NBDC）は、この流れの中で、生命科学データベース横断検索と生命科学系データベースアーカイブ、及びIntegbio データベースカタログを公開し、データの利用者だけでなく開発者の課題に応える活動を行っている。横断検索では今年、ダンプデータ横断検索機構の開発と、検索クエリログの分析に基づく検索結果スコアリングアルゴリズムの改定を実施し、検索対象の拡大と検索結果の適正化を達成した。アーカイブでは寄託件数が130件を突破し、Scientific Dataの推奨レポジトリに認められ、コミュニティにおける基盤の1つとして認知されつつある。

国際社会では研究データの適切な循環のため、FAIR 原則（Findable、Accessible、Interoperable、Reusable）に従うことが標準となりつつある。横断検索とアーカイブでこの条件をサポートし、日本の生命科学界の研究データ循環のメインパイプとなるよう、引き続き改善と品質向上に努めていく。

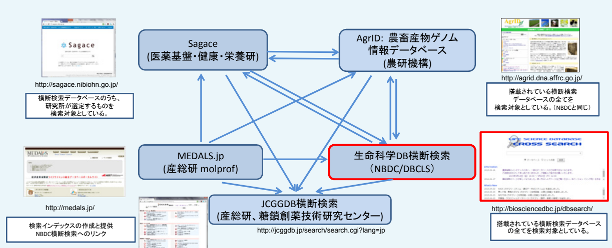
生命科学データベース横断検索とは

■システム概要



キーワードに対し、データベースの一括検索を行い、
ヒット件数と、検索結果（対象へのリンク、ヒット箇所）を返す**Web検索システム**

■組織間連携



4省庁の別組織にサーバを分散配置し、相互に参照しながら検索可能。 → **615件**の生命科学系データベース**6400万件**以上のエントリを対象に検索可能 ※2017年10月時点

検索利用者の意図抽出アルゴリズムの構築

■横断検索のページランクスコア算出方法の改善

横断検索の検索結果の表示順を改善し、利用者が素早く目的の情報にアクセス可能とするため、検索エンジンアプリのElasticsearchにおけるアルゴリズム "TF/IDF practical scoring function of Lucene" のパラメータ設定について検討した。

$$score(q,d) = queryNorm(q) \cdot coord(q,d) \cdot \sum \left(tf(t\ in\ d) \cdot idf(t)^2 \cdot t.getBoost() \cdot norm(t,d) \right) \quad (t\ in\ q)$$

- $score(q,d)$: The relevance score of document **d** for query **q**
- $queryNorm(q)$: The query normalization factor $1 / \sqrt{sumOfSquaredWeights}$
- $coord(q,d)$: The coordination factor (add the rate of matching query words, relates to partial match)
- $tf(t\ in\ d)$: The term frequency for term **t** in **d**
- $idf(t)^2$: The inverse document frequency for **t**
- $t.getBoost()$: The boost that has been applied to the query
- $norm(t,d)$: The field-length norm, combined with the index-time field-level boost, if any
- $(t\ in\ q)$: The sum of the weights for each **t** in **q** for **d**

https://www.elastic.co/guide/en/elasticsearch/guide/2.x/practical-scoring-function.html

→ **動的に設定できるboostパラメータを検索クエリによって変化するよう修正を実施した。**

■ 複数語クエリのパターニングによる検索意図の解釈

Chemical–chemical context pattern
#C and #C
#C versus #C
#C and #C interaction
#C and #C combination
#C plus #C
#C with #C
comparison of #C and #C
interaction between #C and #C
#C oxidase #C
#C dehydrogenase #C
combine #C and #C
#C transporter #C

Chemical–disease context pattern
#D and #C
#C induce #D
#D treatment #C
treatment of #D with #C
#C #D review
#D with #C
#D child #C
#D induce by #C
#D due to #C
#D treatment with #C
role of #C in #D
#C metabolism and #D

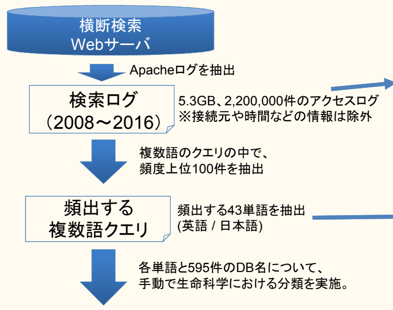
（Huang and Lu, Database, 2016, 1–15, doi: 10.1093/database/baw025）

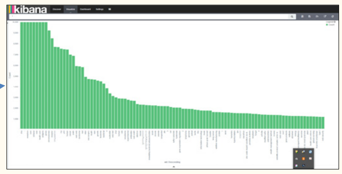
NCBIでは、2015年にPubmedの検索クエリを解析し特に、Chemical（#C）とDisease（#D）を含む複数語検索のペアを抽出した。このペアを分類し、左に示すような頻出パターンに分類した。この情報を元にPubmedの検索結果表示方式への反映を行った。同様に**2単語以上**の複数語検索では、GoogleのVertical search（画像検索・ニュース検索等）に対応するような、特徴的な「サブクエリ」（″画像″や″ニュース″）が含まれていれば、意図の推定が行い易いことが知られている(This concept is inspired by Yamamoto *et. al.*, "Overview of the NTCIR-12 IMine-2 Task", NTCIR 2016)

→ **検索ログを元に複数語クエリの分類と解釈を実施し、横断検索のboostパラメータの変動アルゴリズムを構築した。**

複数検索クエリを利用した意図抽出の例

■ 検索ログデータからの頻出複数語クエリパターン抽出





frequent query(English)	frequent query(Japanese)	Counts	multiple score category
receptor	レセプター	30,622	-
health care	医療	13,414	医療・薬
antigen	抗原	12,174	-
gene	遺伝子	10,564	ゲノム・遺伝子・RNA＞遺伝子・転写産物
chromosome	染色体	9,232	ゲノム・遺伝子・RNA
rna	リボ核酸	7,562	ゲノム・遺伝子・RNA>機能RNA
p53	p53	7,512	-
cancer	癌	7,012	-
carcinoma	癌	5,917	-
eye surgery	眼科手術	4,750	医療・薬
health	健康	4,719	医療・薬
leukemia	白血病	4,530	-
neoplasm	新生物	2,887	-
experimental	実験的	2,785	-
antibody	抗体	2,712	-
anemia	貧血	2,213	医療・薬
hysteresis	履歴効果	2,063	タンパク質>ドメイン・モチーフ・立体構造
lymphoma	リンパ腫	1,950	医療・薬
lang cancer	肺癌	1,652	医療・薬
hypothyroidism	甲状腺機能低下症	1,556	医療・薬
ips	IPS	1,543	細胞・組織
pereskia aculeata	モクキリン	1,514	-
arthritis	関節炎	1,415	医療・薬
cox2	シクロオキシゲナーゼ	1,326	-
epilepsy	てんかん	1,292	医療・薬
alzheimer	アルツハイマー	1,282	医療・薬
velo pharyngeal insufficiency	口蓋帆咽頭の不足（医学用語?）	1,280	医療・薬
sarcoma	肉腫	1,272	-
brachio oto renal syndrome	腎症候群	1,268	医療・薬
health care reform	医療改革	1,246	-
global warming	地球温暖化	1,236	-
ibuprofen enantiomer	イブプロフェン	1,223	医療・薬
health care management	医療管理	1,182	-
hepatitis	肝炎	1,153	医療・薬
prostate	前立腺	1,134	-
antisense	アンチセンス	1,121	ゲノム・遺伝子・RNA
diabetes mellitus	糖尿病	1,102	医療・薬
inborn errors	先天異常	1,100	-
dermatitis	皮膚炎	1,098	医療・薬
air pollution	大気汚染	1,072	-
neuroblastoma	神経芽細胞腫	1,050	医療・薬
tuberculosis	結核	1,046	医療・薬
adenoma	腺腫	1,045	-

生命科学データベース横断検索について以下の改修を実施した。

- 一部の検索クエリパージングに例外処理を追加
- 検索結果の表示に影響するランキングスコアを、指定した検索クエリに該当する場合に通常の50倍になるよう設定（×1～10のスコア変動アルゴリズムは別目的で既に利用されていた）

■ 修正前後の検索結果順の比較

search query	Before the tuning			After the tuning			Difference
	1st result	2nd result	3rd result	1st result	2nd result	3rd result	
toxic データベース	医学・薬学予稿集全文データベース "毒性作用機序データベースの開発"	医学・薬学予稿集全文データベース "毒性データベースと毒作用予測（2）：データベースの利用"	医学・薬学予稿集全文データベース "毒性データベースと毒作用予測（1）：データベースの構築"	Integbio database catalog "中毒情報データベース"	Integbio database catalog "Open TG-GATEs"	Integbio database catalog "TOXNET"	"データベース" multiplied the score of database catalog score.
結核　症状	J-STAGE "原発性鼻腔結核の1症例 A case of primary nasal tuberculosis"	J-STAGE "喉頭結核の3症例 Three cases of laryngeal tuberculosis"	"医学・薬学予稿集全文データベース" Mycobacterium tuberculosis感染腹腔マクロファージの壊死の阻害におけるカスパーゼ9の役割	"PDQ日本語版" PDQ日本語版（患者様向け）：用語：結核	"UMIN-CTR(臨床試験データベース)" UMIN000010588：結核および肺Mycobacterium avium complex症治療薬の薬物相互作用に関する臨床症状と検査値からの検討	"UMIN-CTR(臨床試験データベース)" UMIN000016454：結核菌感染症における鉄代謝関連分子ヘプシジン（hepcidin）の治療反応性のバイオマーカーとしての有用性を明らかにする観察研究	"結核" multiplied the score of "medical/clinical category" score.
gene enolase	"医学・薬学予稿集全文データベース" ラット卵巣で発現するエノラーゼ2	"医学・薬学予稿集全文データベース" α-エノラーゼ変異体とc-myc P2プロモーターとの親和性解析	"医学・薬学予稿集全文データベース" 麹菌のエノラーゼプロモーター（PenoA）へのシスエレメント導入効果	"医学・薬学予稿集全文データベース" ラット卵巣で発現するエノラーゼ2	"医学・薬学予稿集全文データベース" α-エノラーゼ変異体とc-myc P2プロモーターとの親和性解析	"医学・薬学予稿集全文データベース" 麹菌のエノラーゼプロモーター（PenoA）へのシスエレメント導入効果	"gene" multiplied the score of "gene/transcript" category. But scores of documents are bigger than the multiplied scores.

→ **検索語によって、明らかに改善されたと判断できるケースと、変化の無かったケースが観察された** → **2017年5月、本番環境に反映**

生命科学系データベースアーカイブとは

■サービス概要

生命科学系データベースアーカイブは、国内のライフサイエンス研究者が生み出したデータセットをわが国の公共財としてまとめて長期間安定に維持保管し、データ説明（メタデータ）を統一して検索 を容易にすると共に、利用許諾条件などの明示を行うことで、多くの人が容易にデータへアクセスしダウンロードを行えるようにしている。データを長期にわたり保全し、データベース作成者のクレジットを明示する一方、公的機関や民間等様々なユーザが利用しやすい形にすることで、それぞれの研究の生命科学へのいっそうの貢献を支援する。当アーカイブは2017年度、Nature publishingのScientific Data誌 の推奨レポジトリに認定された。

■主な特長

- NBDC「統合化推進プログラム」で産生されたデータベースも含め、130以上のデータベースをダウンロード可能
- 各データベースのメタデータが付与されており、データの再利用が容易
- 各データベースの利用許諾は、明確かつシンプル（CC BY-SA ライセンス等）
- 各データベース及びそのサブセットにはDOI (Digital Object Identifier)が付与されており、論文等で簡単に同定・引用できる
- 多くのデータがRDF化対応済み
- SCIENTIFIC DATA 推奨レポジトリに認定されている

データベースの寄託を募集しています。お気軽にご相談ください。

まとめ・今後の方針

- 生命科学データベース横断検索では、検索ログを利用した意図抽出アルゴリズムを設定し、より利用者がデータを発見し易い基盤となるよう、改善を継続している。
- 生命科学系データベースアーカイブでは、寄託データを長期にわたり保持する活動をおこない、Scientific Data誌の推奨レポジトリに認定され、利用者の継続的なアクセスに貢献している。
- 国際社会では研究データの適切な循環のため、FAIR 原則（Findable、Accessible、Interoperable、Reusable）に従うことが標準となりつつある。横断検索とアーカイブでこの条件をサポートし、日本の生命科学界の研究データ循環のメインパイプとなるよう、引き続き改善と品質向上に努めていく。

トーゴーの日シンポジウム2017 ～バイオデータベース：「作る」から「使う」へ～
会期：2017年10月4日(水)－5日(木)　於：東京大学弥生講堂