生命科学データベース横断検索の継続的な発展と展望

○大波純一[1]、杉崎太一朗[2]、友田史緒里[2]、牧口大旭[2]、川本祥子[3,4]、畠中秀樹[1]

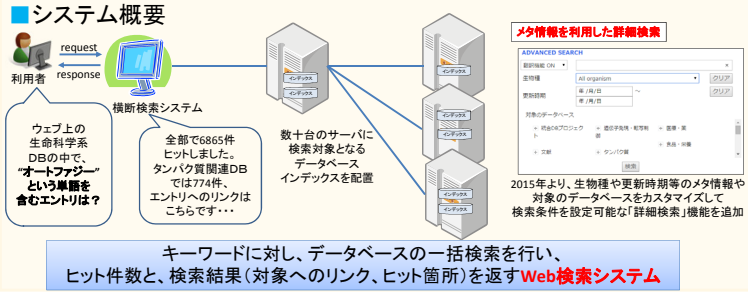
1.国立研究開発法人科学技術振興機構バイオサイエンスデータベースセンター（NBDC）、2.三井情報株式会社、
3.大学共同利用法人情報・システム研究機構ライフサイエンス統合データベースセンター（DBCLS）、4.情報・システム研究機構 国立遺伝学研究所

要旨

「散在するデータを発見しやすくする」ための情報検索基盤の価値は、近年のデータ駆動型研究や機械学習を利用した研究の拡大、オープンデータ推進の風潮に伴い、さらに高まりつつある。生命科学分野の研究データに特化した検索基盤である生命科学データベース横断検索は、今年で公開から約10年を迎えた。利用者から頂いた貴重なご意見や連携機関による多大なご協力の元、対象データベースの数も640件以上に増加し、約1億件のエントリにアクセスできる大規模なサービスに成長した。この一方で、利用者の需要や扱うデータの質も徐々に様変わりしている。例えば検索結果のリッチスニペットへの需要が増加し、NGSによるものやハイスループットな研究結果が増えるにつれ各データベースの構造も複雑になり、規模も大きくなってきている。加速的に変化し続ける生命科学データの検索基盤としての、これまでの実績と将来の展望について議論を行う。

生命科学データベース横断検索とは

システム概要

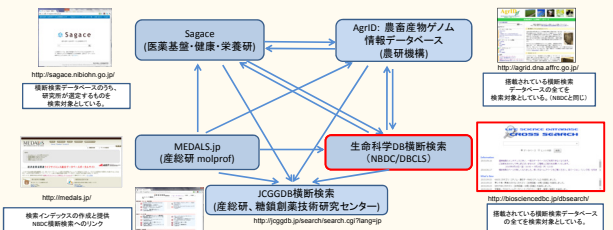


キーワードに対し、データベースの一括検索を行い、
ヒット件数と、検索結果（対象へのリンク、ヒット箇所）を返すWeb検索システム

沿革

2008年	大学共同利用法人　情報・システム研究機構（ROIS） ライフサイエンス統合データベースセンター（DBCLS）にて 文部科学省「統合データベースプロジェクト」のサポートによりサービス開発開始。
2009年1月	独立行政法人産業技術総合研究所（産総研）生命情報工学研究センターの協力で 糖鎖関連のDB（7件）を拡充し、検索対象DB数が232件に拡大。
2010年5月	産総研バイオメディシナル情報研究センター（BIRC、後に創薬分子プロファイリング研究センターへ引継）の 「MEDALS横断検索」のDB（21件）と相互連携。検索対象DB数：253件に。
2011年4月	JSTバイオサイエンスデータベースセンター（NBDC）へ移管。
2011年9月	英語版を提供開始。
2012年4月	農業生物資源研究所、医薬基盤研究所との連携を開始。
2014年9月	DBCLS組織移転。DBCLS設置のサーバを、JST共通IT基盤およびNBDC札幌データセンタへ移設。
2015年10月	検索エンジンをHyper estraierからElasticsearchへ変更
2016年1月	NBDC札幌データセンタのサーバをJST共通IT基盤へ移設
2017年5月	検索クエリ処理アルゴリズム刷新
2019年5月	JST共通IT基盤刷新予定

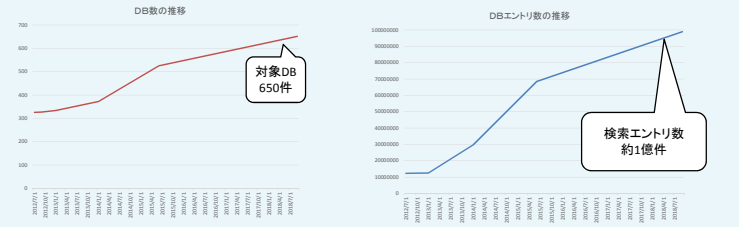
組織間連携



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。

650件の生命科学系データベース
約1億件のエントリを対象に
検索可能　※2018年10月時点

対象データの増加



インターフェースの変遷

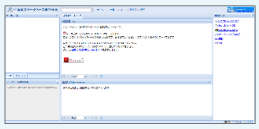
横断検索のトップページの10年

横断検索のインターフェースは、時代の要請やヒューマンインターフェースの検討を元に改修を継続してきた。

2008

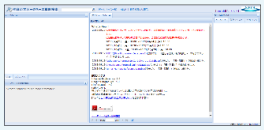
最初期のインターフェース。
ディレクトリ型検索寄りの
見かけとなっている。

2010



絞込みの際に有用なツリー型のカテゴリ表示や、検索クエリから関連語句をサジェストするツール表示等の多機能を実装し、フレームに分割して表示させるようにした。

2012



サジェスト情報や関連遺伝子情報をタブ付のフレームにまとめ、表示を刷新した。現在も「旧インターフェース」として利用可能。

2015

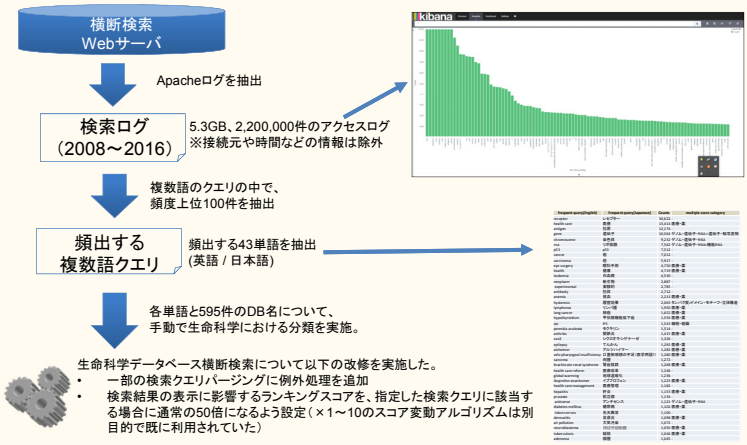
ファーストインプレッションにおける視覚情報を最小化させるため、Googleライクな画面とした。先行していた医薬基盤研Sagaceの画面を参考として構築した。

2018

検索例の表示や、検索履歴をクッキーで保存させる機能の追加を実施。連携機関との協調体制を示すため、Integbioポータルサイトの色調に合わせた表示とした。

検索アルゴリズムの強化

検索ログデータからの頻出複数語クエリパターン抽出



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。

- 一部の検索クエリイン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

本番環境に反映

まとめ

- 2018年、生命科学データベース横断検索は構築されてから10年が経過した。
- 検索対象データベースの追加、連携機関との協調を継続し、長期的な進化を達成している。
- インターフェースの刷新や検索アルゴリズムの変更等、利用者目線での改善も継続している。
- 引き続き、次の10年に向けて改善や強化を継続していく。

トーゴーの日シンポジウム2018 ～バイオデータベース：つないで使う～
会期：2018年10月5日(金) 於：日本科学未来館