生命科学データベース横断検索のメタ情報利用とインフラ環境強化

○大波純一[1]、杉崎太一朗[2]、青木健一[2]、平井信一[2]、牧口大旭[2]、川本祥子[3]、畠中秀樹[1]、三橋信孝[1]

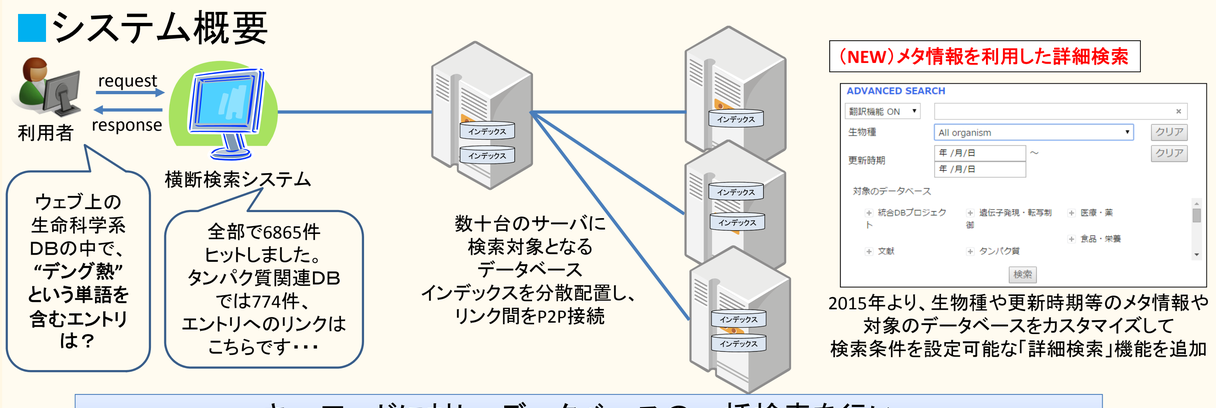
1.国立研究開発法人科学技術振興機構バイオサイエンスデータベースセンター（NBDC）、2.三井情報株式会社、
3.大学共同利用法人情報・システム研究機構ライフサイエンス統合データベースセンター（DBCLS）

要旨

Web 上に公開されている多様な生命科学系公開データベースを、深層ウェブのエントリも含めて一律にテキストー致検索を行うための基盤として、「生命科学データベース横断検索」を公開している。現環境はライフサイエンス統合データベースセンターの協力の元、医薬基盤・健康・栄養研究所、農業生物資源研究所、産業技術総合研究所・創薬分子プロファイリング研究センター、産業技術総合研究所・糖鎖創薬技術研究センターが協調してインフラ整備とデータ更新を実施する体制となっている。利用者はここから533 件のデータベース、68,549,042 件のエントリ(2015 年7 月時点)に直接アクセスすることができ、研究テーマに絞った情報の抽出や先行研究の状況調査を効果的に行うことができる。また、各エントリから抽出されたメタ情報を利用するため、キーワード以外の検索条件(生物種や更新時期)を指定できる詳細検索機能の公開を開始した。この一方で、対象のデータベースの数は3 年間で1.6 倍、エントリの数は5.5 倍に増え、検索速度等のパフォーマンス低下が指摘されていた。そこで2014 年後半から、ハード環境の刷新と検索エンジンの刷新を実施している。この大規模な検索基盤の現在と未来について報告する。

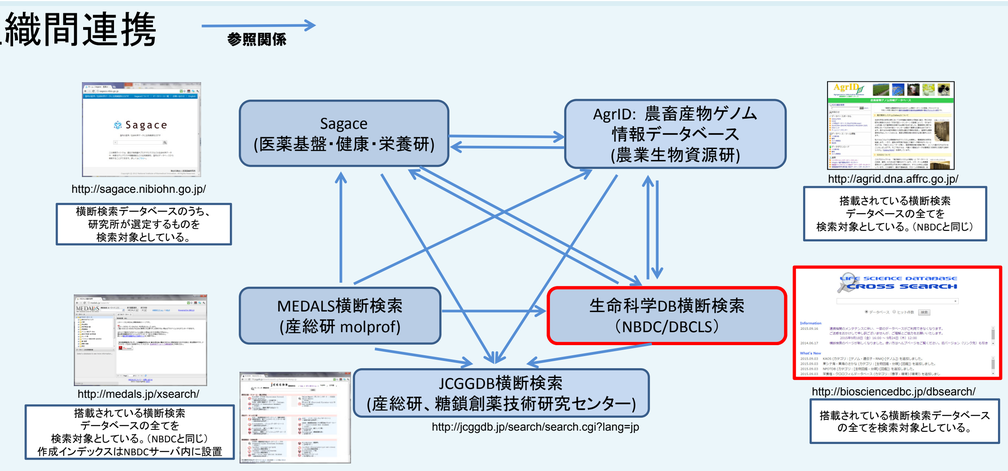
生命科学データベース横断検索とは

■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**

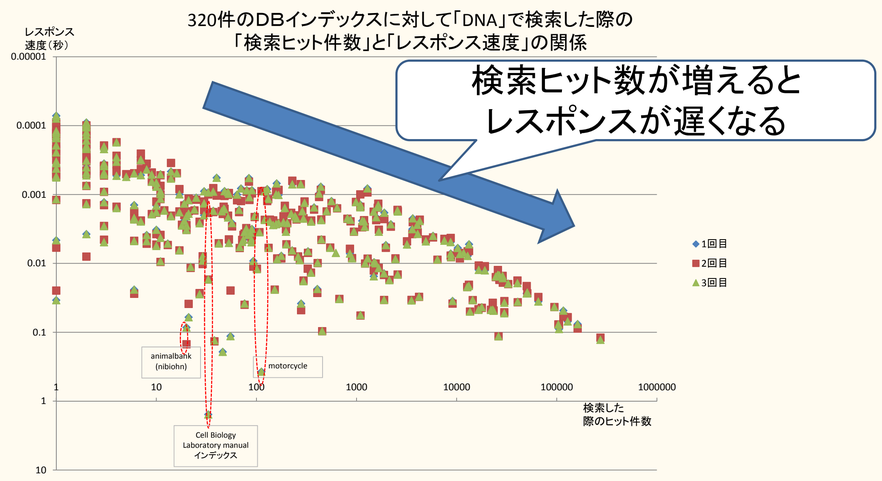
■組織間連携

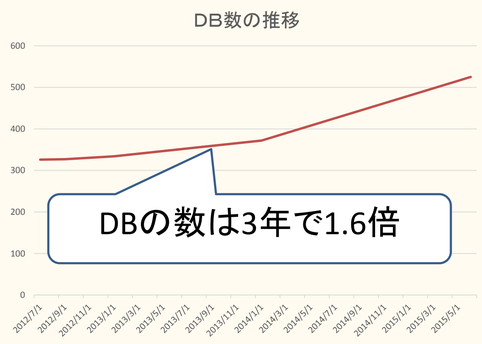


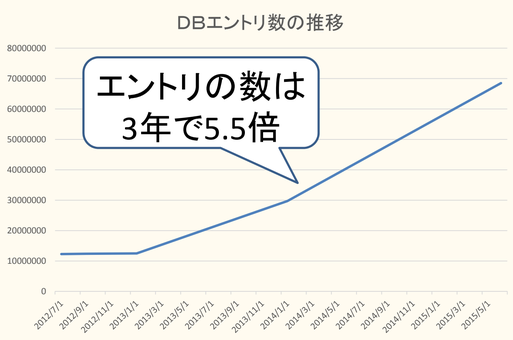
4省庁の別組織にサーバを分散配置し、相互に参照しながら検索可能。

→ **533件**の生命科学系データベース **6854万件**以上のエントリを対象に検索可能　※2015年7月時点

性能強化の必要性







パフォーマンスチューニングを実施しながら検索対象を追加しないと、パフォーマンスは落ちる一方

ハードウェア環境の増強

2014年10月、サーバのリプレースに合わせ、ハードウェア環境の強化を実施した。

@DBCLS旧サーバ室 旧検索サーバ(2008年～2014年)	@JST東京本部共通IT基盤 新検索サーバ(2014年～)
・CPU: Intel Xeon 64bit (2.6GHz, 4core) ・メモリ: 16GB ・ディスク: 89GB ・最大通信同時接続数：256件 ×45台	・CPU: Intel Xeon 64bit (2.6GHz, 4core) ・メモリ: **29GB** ・ディスク: **292GB** ・最大通信同時接続数：**無制限** ×50台

過去の検証（データ未公開）で、横断検索のパフォーマンスのボトルネックとなっていた、メモリと通信同時接続数を大幅に増強し、検索対象DB増加後も過去と同じパフォーマンス（毎回7秒程度）を維持することができた。

→ **ハード環境の増強で、データ量や通信量の伸びに対応**

検索エンジンアプリケーションの最適化

横断検索サービス公開（2007年）から使われている検索エンジンアプリケーションHyper Estraier (以後HEと呼称。http://fallabs.com/hyperestraier/)には、以下の懸念が存在した。

1.個人開発のアプリケーション
課題が発生した場合、企業のような十分なサポートが得られない可能性がある

2.セキュリティリスク
2007年以降アプリケーションの更新が停止しており、将来、構成モジュールの最新化や別OSへの対応時にセキュリティリスクが生じる可能性がある

3.新機能や環境への対応
近年のビッグデータ利用のトレンドに伴い、後発の検索エンジンではMap reduce技術やパフォーマンスチューニングの洗練化などの技術が発達している

このため、新興の検索エンジンアプリケーションの中で、上記懸念をクリアするElasticsearch（以後ESと呼称。https://www.elastic.co/jp/）を選定し、テストを行った。

検索条件：　現在横断検索に登録されているデータの一部、51,768,997 件の文書（94.7GB）を対象とし、HEでは10件の検索結果を取得、ESでは0・10・25・50・100・250・500・1000件の検索結果を取得し、最長の検索時間5回分の平均値を計測した。

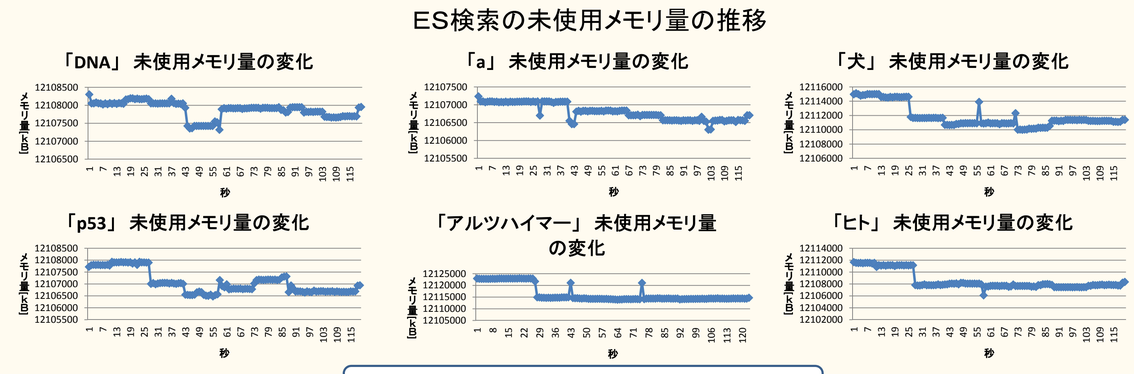
検索時間の5回平均 (単位: 秒)

	DNA	p53	a	アルツハイマー	犬	ヒト
Hyper estraier	20.24	10.884	7.278	3.509	3.004	3.356
Elasticsearch	0.174	0.263	0.223	1.164	0.641	0.445

検索時間は3.01倍から116.32倍高速化

次に同じ検索対象に対し15秒おきに6種類のクエリで検索を行い、1秒おきに未使用メモリ量の測定を行った。

ES検索の未使用メモリ量の推移



ESはHEの約1/5のメモリ使用量

→ **Elasticsearchでは高速化とメモリ使用量の効率化が見込める**

まとめ・今後の方針

・生命科学データベース横断検索の検索対象増大に伴い、パフォーマンス低下が懸念され、ハードの刷新を行って検索時間の現状維持に繋げた。
・検索エンジンの課題やパフォーマンス低下に対応する新アプリケーションとして新たにElasticsearchを選定し、その動作検証を行った。
今後Elasticsearchへの刷新（12月公開予定）を行い、より高速で効率的な検索環境を提供する。

トーゴーの日シンポジウム 2015 ～統合データベースとオープンライフサイエンス～
会期：2015年10月5日(月)－6日(火)　於：東京大学弥生講堂

Licensed under a Creative Commons 表示 2.1 日本 license
© 2015 バイオサイエンスデータベースセンター