National Bioscience Database Center

独立行政法人 科学技術振興機構 Japan Science and Technology Agency

生命科学系データベース横断検索サービスの現状と今後

○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]、奥村利幸[2]、川本祥子[3]、畠中秀樹[1]、三橋信孝[1]

1.独立行政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）

現在の生命科学データベース横断検索システム概観

■検索インターフェース

http://biosciencedbc.jp/dbsearch/ もしくは「生命科学横断検索」でWeb検索

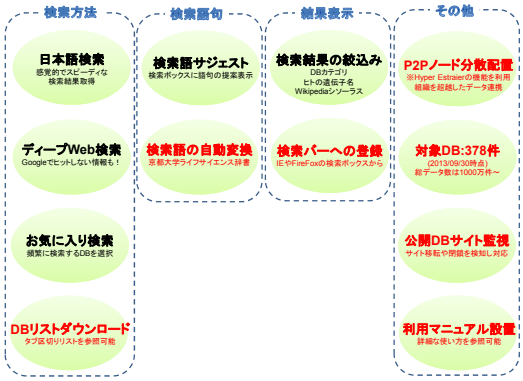
■英語版

http://biosciencedbc.jp/dbsearch/?lang=en

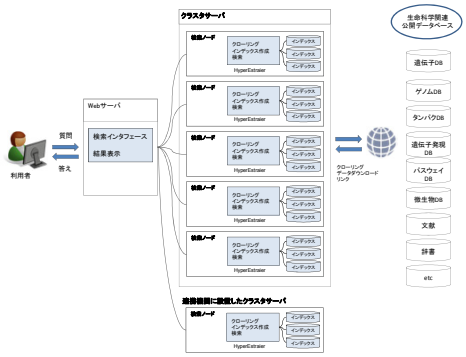
■モバイル版

http://biosciencedbc.jp/dbsearch/mobile/

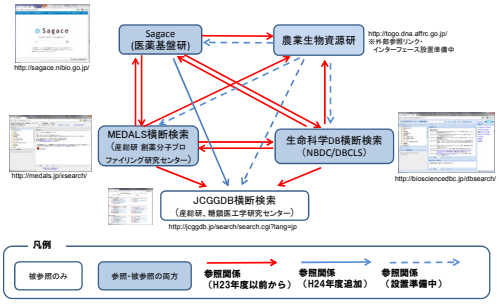
■機能と特徴 （赤字：機能更新・追加）



■システム構成



■クラスタサーバの組織間での相互乗り入れ図

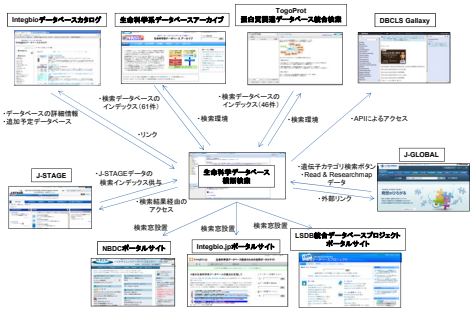


統合的な検索に向け

1. 幅広い分野のデータベースの相互接続

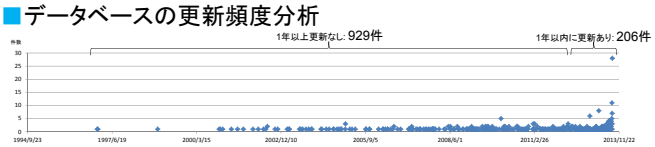
2. メタ情報統合化の基準を統合DBプロジェクト内で検討

■外部サービスとの連携



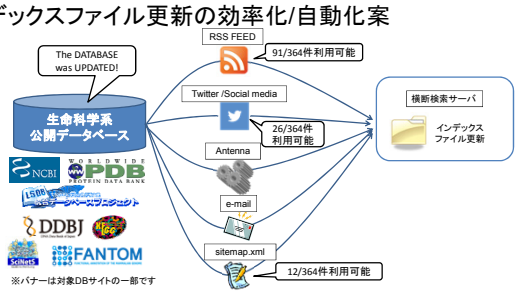
これからのシステム機能強化方針

■データベースの更新頻度分析



Integbioデータベースカタログ（ポスター番号1を参照）で公開されている、稼働中のライフサイエンス系データベース情報1135件を利用し、2013年9月2日時点の、DBサイトトップページの最終更新日を確認した。この結果、最終更新日が直近1年以内のデータベースサイトは206件（約18%）に留まっていた。定期的に公開DBのデータを取得し、インデックスファイルを更新する横断検索システムでは、より効率的なデータ更新の確認と取得が必要と考えられる。

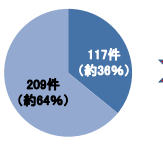
■インデックスファイル更新の効率化/自動化案



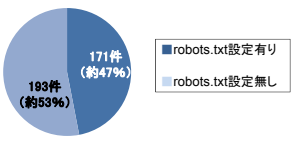
※バナーは対象DBサイトの一部です

■robots.txtを公開しているDBサイトの割合

A. 2012年9月時点での生命科学データベース横断検索における robots.txt設定データベースの割合



B. 2013年10月1日時点の生命科学データベース横断検索における robots.txt設定データベースの割合



■そして理想的な統合検索へ・・・

【現状】人間によるデータベースの取得範囲の調査・取得作業

無精造なデータベースでも直接アクセスし、要素を見分ける

①全データの件数は？
②各データのURLは何と何と何？
③どのような方法で全データをダウンロードできる？
④サイトが利用規約や検索エンジン宛の通達でダウンロードについて許可している？
⑤英語版のデータと日本語版のデータはどのように分かれている？
⑥ダウンロードの時に相手のサーバに負荷をかけて嫌がらせしまわないか？
⑦各データの中の検索範囲はどうするか（著作権表記や宣伝は検索対象にしない）
⑧データ内容に更新があったらどのようにして分かるか？最終更新日はいつか？
⑨現在の横断検索に搭載してパフォーマンスに問題はないか？
⑩各データがオリジナルではなく、他のデータベースのコピーではないか？
⑪過去に横断検索に搭載した類似データベースではどのように扱っているか？
⑫ページ内で生命科学系データとして検索するべき範囲はどこか。
⑬遺伝子名は他のDBサイトと共通でLinked Dataとして利用できるか。
⑭生物種は学名か一般名か？
⑮構造データやバインディングモチーフ、タンパク質ファミリーなどのリンクは可能か。
⑯配列データはどのようなフォーマットで利用可能か。
etc…

取得するデータの例（赤枠：検索インデックスに登録する範囲）

Entry	5422 CDS T01001
Gene name	POLA1, POLA, p180
Definition	polymerase (DNA directed), alpha 1, catalytic subunit (EC:2.7.7.7)
Ortholigy	K02320 · DNA polymerase alpha subunit A [EC:2.7.7.7]
Organism	hsa Homo sapiens (human)
Pathway	hsa00230 · Purine metabolism
Module	hsa_M00261 DNA polymerase alpha / primase complex
Class	Metabolism; Nucleotide metabolism; Purine metabolism [PATH:hsa00230] Metabolism; Nucleotide metabolism; Pyrimidine metabolism [PATH:hsa00240] Genetic Information Processing; Replication and repair; DNA replication [PATH:hsa03030] BRITE hierarchy
Motif	Pfam: DNA_pol_B DNA_pol_B_exo1 zf-DNA_Pol DNA_pol_alpha_N Motif
Other DBs	NCBI-GI: 108507301 NCBI-GeneID: 5422 OMIM: 312040 HGNC: 9173 HPRD: 02416 Ensembl: ENSG00000101868 Vega: OTTHUMG00000021277 UniProt: P09884
Position	Xp22.1-p21.3
AA seq	MAPVHGDDSLSDSGSFVSSRARREKKSKKGRQEALERLKKAKAGEKYKYEVEDFTGVYEE VDEEGYSKLVQARQDDDWVDDDGIGYVEDGREIFDDDLEDDALDADEKGKDGKARNKDK RNVKKLAVTKPNNKSMFIACAGKKTADKAVDLSKDGLLGDILQDLNTETPQITPPPVME LKKKRRSQIASPNRFSVHTATAVPSQKDASPVSRKEPPLTPVPLRAEFAGDDDVQVESTEE EQESGAMEFEDGDFDFEPMEVEEVDLEPMAAKAWDKESEPAEEVKQEADSGKGTVSYLGSF LPDVSCWDIDGEGDSSPSVQEVQVDSSHLPLVKGADEEQVFHFYWLDAYEDGYNQPQVVF LFGKVWESAETHVSGCDVMVKNERTLYFLPREMKDLNTGKETGTPRSMKDVYEEFDEK IATKYKMKFKSKPVEKNYAFEIPDVPEKSEYLEVKYYSAEMPQLPQDLKGETFSHVFGTN

現在横断検索では、人手で最新の情報を確認し、公開データベースのエントリーごとに重要な要素を見分けて検索インデックスに登録している。このため、一般的なWeb検索と比べ、高い網羅性と精度で生命科学系データにアクセスすることができる。但しこの手法では、統合的な検索（DB内項目の意味を理解した検索方法）とならない上、情報確認に時間がかかる。

【将来】データベースの構造化・公開データベースごとのRDFメタ情報スキーマの埋め込み

生命科学系データの統合的な利用
ユニバーサルで負担の少ない取得機構

RDF化・構造化された公開データベース

今後は、統合的な検索の実現のため、構造化（RDF化など）されたデータベースへの取得方法の検討や、生命科学分野の研究に有用な統合項目の検討、データベース取得機構のユニバーサル化などを実施し、情報確認の効率化を推進する。また、連携機関、検索対象データベース管理者、利用者からの意見も積極的に取り入れて、更なる改善に努める方針である。

トーゴーの日シンポジウム2013　ここまできたライフサイエンスデータベースの統合

会期：2013年　10月4日（金）・5日（土）　於：時事通信ホール