生命科学横断検索サービスの開発

http://lifesciencedb.jp/dbsearch/

〇牧口大旭1)，高橋順子1)，川原麗子1)，谷口仁2)，奥村利幸1)，横山貴央3)，川本祥子3)

1)三井情報(株)，2)日立ソフト(株)
3)ライフサイエンス統合データベースセンター

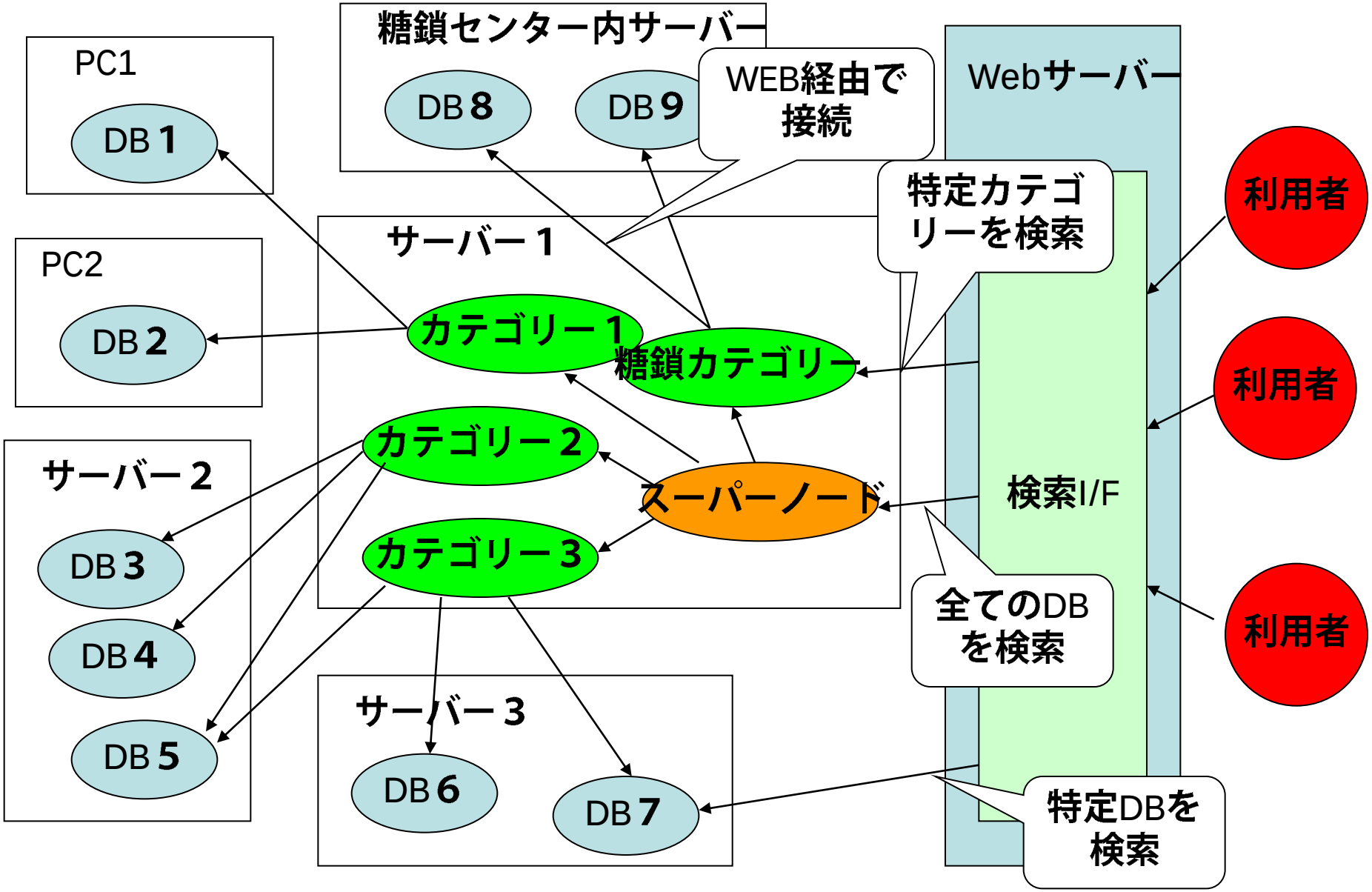
生命科学情報横断検索サービスとは？

- 国内外の有用な生命情報データベースと、特許や雑誌・新聞記事、総説等の文献情報などを一括で検索できる、網羅的なライフサイエンス情報検索サービスです。
- 現在２００以上データベース群を収録している他日本国特許、蛋白質核酸酵素(共立出版)、新聞のライフサイエンスに関わる過去記事(科学新聞)、文科省のゲノムに関わる主要プロジェクトの報告書等のオリジナルコンテンツも収録しています。

システムの構成（データベース）

- 検索エンジンに**HyperEstraier**を用いています。
- HyperEstraierの持つP2P機構を利用して柔軟な構造と高いスケーラビティをもつ検索システムを低コストに実現可能です。
- 検索用のDB本体にあたるインデックス情報をもつ実体ノード(次項DB)を任意のサーバー/PCに配置し、それらをある単位で取りまとめる、リンク情報のみを持つ参照ノード(次項カテゴリ/スーパーノード)により構成されています。
- 各ノード間はHTTP1.0によるP2Pで接続されており、物理的に離れていても容易に接続が可能です。また、リンク情報の組換えることで、既存のリソースを利用しつつ、様々なバリュエーションをもった検索システムを提供できます。

システム構成（データベース）



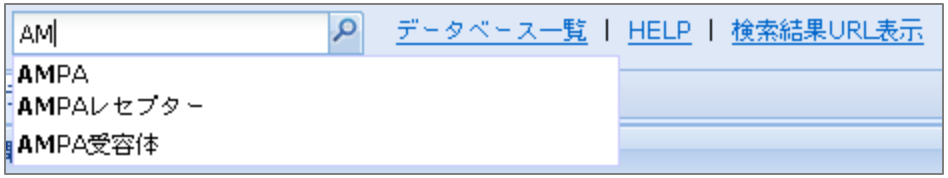
システム構成（メリットとデメリット）

- P2P機構を用いたシステムのためデータベースの拡張の際には接続先のノードを置いたPC/サーバーを増やし、ネットワークに接続させるだけで対応できます。あらかじめ全体のデータベース規模を見積もる必要が無く、状況に応じた無駄の無い対応が可能です
- またスーパーコンピュータ等大規模なリソースは必要なく、再利用目的で古いPC等でもシステムに組み込むことが可能で、非常にコストパフォーマンスに優れています
- ただし、管理対象が分散してしまい、ある程度の規模になると別途これらを統合管理するためのシステムが必要になります
- 前項で述べたデメリットを解消すべく、管理用のシステムも併せて開発しています

検索インターフェース 全体図

検索インターフェース　①

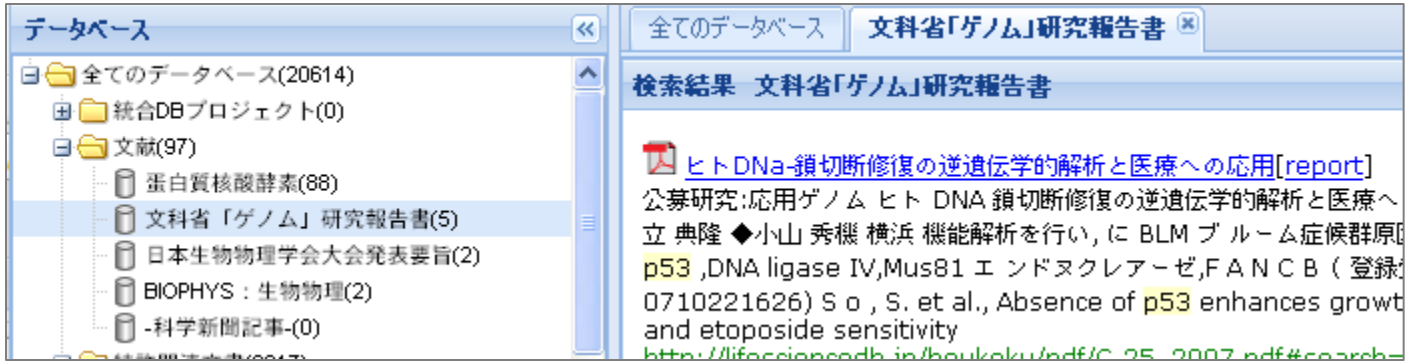
- 検索クエリの入力
  - 全ての検索クエリは日英双方向に翻訳され検索されます（翻訳辞書にLSD、形態素解析にMeCabを利用）
  - また簡易的なサジェスト機能を提供しています



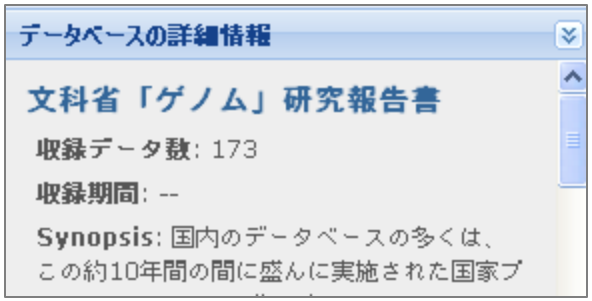
  - 空白にてAND検索｜にてOR検索！にてANDNOT検索等を利用できます
  - クエリボックス横の各リンクはDBの一覧情報や、検索結果を直接表示するためのURLの発行などの機能を提供しています

検索インターフェース②③④

- **検索結果の絞込み**
  - ②のツリーパネルでツリーノードを選択することで検索結果をカテゴリーやDB毎に絞り込めます。絞り込んだ結果は③の検索結果表示部にタブとして追加されます。
  - 名称の右側の数字は検索結果のヒット件数です。

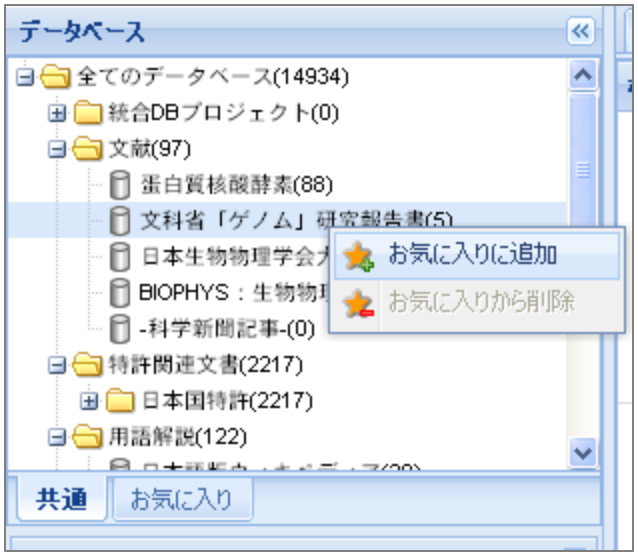


  - 選択中のデータベースは詳細情報が④に表示されます。



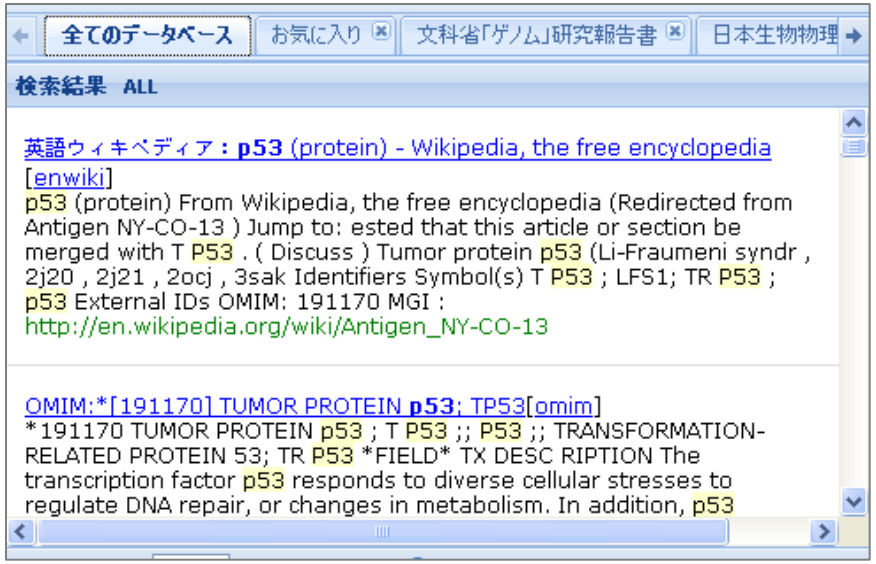
検索インターフェース　②④

- **お気に入り機能**
  - ②のツリー上で、右クリックにて表示されるメニューよりデータベースをお気に入りに追加できます
  - お気に入りに追加されたデータベースは、ツリーパネルのお気に入りタブにて確認できます
  - お気に入りタブを開くことで登録されたデータベースのみの検索結果を表示できます



検索インターフェース　③⑤

- 検索結果は③に表示されます。カテゴリーや絞込みの結果はタブが追加され検索結果が表示されます
- 全てのデータベースタブを表示している時には⑤の遺伝子リストが表示されます
- 遺伝子リストにはクエリに対応した遺伝子の一覧が表示され遺伝子名をクリックすることで該当遺伝子に関するレコードの検索ができます



検索インターフェース⑥

- 外部リンク
  - ⑥は外部へのリンクリストになっています
  - NCBI、PubMed、Googleに対しては、入力された検索クエリを用いて直接検索します。日英双方向の翻訳機能が適用されます

検索APIの提供

- 検索結果をJSONもしくはJSONPにて受け取れるAPIを用意しています。
  - http://lifesciencedb.jp/dbsearch/に対し以下のパラメータをGETクエリで投げてください
    - phrase:検索フレーズ
    - nodeid:検索対象ノードに対するIDです。カテゴリーやDB毎で結果を取得する際に必要になります。省略した場合は全データベースの検索結果が返ります
    - max:検索結果取得件数です。デフォルトは上位50件です
    - outputStyle:JSONもしくはJSONPを指定してください。省略した場合は、検索インターフェースに結果が反映されたHTMLが返ります
    - callback:JSONPを指定した場合のcallcack関数を指定してください

EX:http://lifesciencedb.jp/dbsearch/?phrase=p53&nodeid=report&outputStyle=JSON