セマンティック検索ポータル

DayPort（ベータ版）のご紹介

Daylight
Chemical
Information
Systems, Inc.

Daylight社製品は、創薬研究支援のための強力なシステム構築ツール群です。お客様のニーズに併せて、多機能で高速な化合物情報処理システムを柔軟に構築できます。今回は、Daylight社にて現在開発中のセマンティック検索ポータルDayPortを紹介します。

■DayPortシステム

現在開発中のDayPortは、セマンティック検索技術を利用したWebポータルです。SMILES/SMARTSの大きな特長である高速な構造検索だけでなく、生物関連情報・論文情報や実験情報など研究に関する様々な情報を検索・管理できます。

■セマンティックな検索

DayPortではセマンティック検索という先進技術を採用することで、文章構成や意味づけを意識した高度な検索を可能にしています。自動的に文章を解釈できるようにするために、DayPortのデータベース（DB）では、データの性質や概念を表すメタデータとその概念を階層化したオントロジーによりデータ間の関係性が明確に定義されます。データの関係性は「主語－述語－目的語」の結合ルールとして記載されます。この結合ルールによって「包含・継承」「化学反応・代謝パスウェイ」「実験プロトコル」など様々な関係性がすべて表現できます。

例えば、化合物「ベンゼン」の記述子情報を扱う場合を考えてみると、従来のDBでは化合物IDカラムといろいろな記述子カラムをもつテーブルを用意して、1行のデータ対として記録していました。これに対してDayPortでは、「ベンゼン（主語）が、記述子としてXXという値（目的語）を、持つ（述語）」という関係として記録します。他にも「ある遺伝子配列（主語）が、とある研究論文（目的語）で、言及されている（述語）」などの記載も、テーブル構成を変更することなく可能になります。

従来のDBでは重要なデータが大量のデータに埋もれてしまいがちでした。これとは逆にDayPortでは、データ同士をつながりやすく加工した小さな"断片"として扱います。これにより、いろいろな視点（関係性）から柔軟にデータを結びつける"データのるつぼ"として活用します。

■MesmirからDayPortへ

これまでDaylight社ではMerlin DBシステムを利用するMesmirという化合物探索ツールの開発が進められていました。DayPortではMesmirがもっていた探索的（芋づる式）な検索をさらに進化させています。あいまいな検索条件からでも対象化合物を柔軟に特定していくというMesmirの設計思想をしっかり受け継ぐだけでなく、検索対象を化合物周辺のDBコンテンツへ拡張しています。

■対象データへすばやくアクセス

オントロジーを活用してデータ間の関係を統制することで、セマンティックという特性を生かしたデータ閲覧が可能になります。DayPortではデータ閲覧のために折りたたみリスト表示・ネットワーク表示ツールが用意されています。

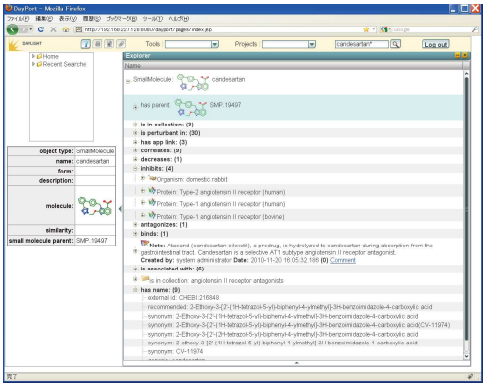

図1: リスト表示ツール

単なるキーワード検索にはない柔軟性も特長の一つです。キーワード検索では検索結果を確認しながら、目的の情報を効率よく抽出できるような検索クエリを類推する必要がありました。このため、DBの構成を意識しつつ、頭のなかで結果を予想してはクエリ作成を繰り返すという婉曲的な作業になりがちでした。DayPortでは、検索結果自体をヒントとして確認しながら関係性を辿ることで、検索条件を微調整できます。実際には条件を調整していることを意識せずに検索することになるかもしれませんが、周辺のデータから目的の情報へすばやくアクセスできます。

また、従来のDBのように検索結果の詳細を確認することももちろん可能です。このほか、検索クエリを保存して再利用できるポータル機能やプロジェクト管理機能も用意されています。

DayPortは自分だけでなく他者の知見も結びつけて迅速な意思決定や意思形成を支援します。

■周辺データの探索

DayPortのネットワーク表示ツールでは対象データの周辺情報を表示します。表示させる関係性の述語をチェックボックスで指定することで、検索条件を調整します。

周辺情報を俯瞰しながらデータを辿ったり、複数のネットワーク表示ツールを起動してそれぞれの関係性ネットワークのパターンを比較できます。

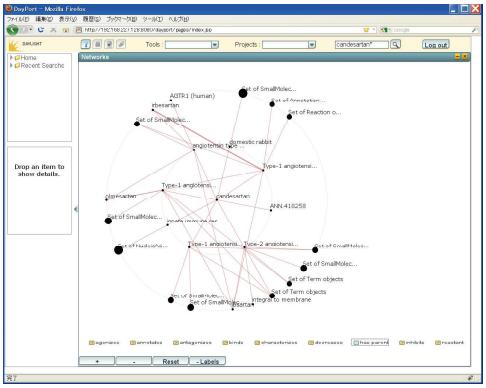

図2: ネットワーク表示ツール

■セマンティックなクエリ作成

DayPortではDB構成を熟知していなくてもクエリメニューから直感的に操作してデータを関連付けます。関連付けが可能な項目がメニュー上にアイコンとともに表示されますので、これら選択しながらクエリを作成します。また、すべてのデータの関係性が「主語ー述語ー目的語」のルールで定義されますので、クエリ内容を確認すると（英語の）文章になっていることも分かります。例えば、図のようなProtocol→Experiment→Proteinと3つの項目を組み合わせでクエリを作成すると、"... is protocol for experiment, and experiment has component protein, ..." のようにクエリが表示されます。

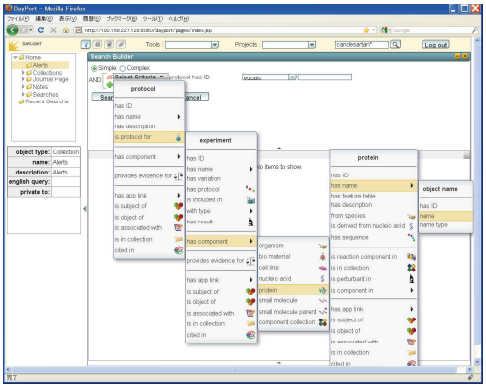

図3: クエリ作成メニュー

■アノテーションによるデータ登録

柔軟性の高いデータも、DayPortでは語彙が統制されることで効率的に登録できます。データを文章の構成要素のように管理しますので、文献からの情報は効果的に抽出できます。例えば、受動態のような目的語と主語が逆転している表現なども正規化して登録します。登録の際にはアノテーションツールを利用します。公開されている論文情報（アブストラクトなど）をアノテーションツールへ入力すると、キーワードを抽出することができます。この際、結合ルールに対応する主語・述語・目的語をキーワードとして認識します。その他、根拠となる情報や添付文書、コメントなども登録できます。さらに、研究には不可欠な信頼性についての情報も登録データに付与できます。信頼性が低い場合は、根拠が不明で検証が必要な仮説として扱います。

DayPortでは組織内外で得られたセマンティックデータを結びつけて知見を育成します。従来のDBでは、テーブル構成が固定されているため、登録データに付随するカラム情報（メタデータ）は一定でした。そのため、広範な情報へ柔軟に対応するためには、テーブルに多くのカラムをあらかじめ持たせる必要がありました。一方DayPortではデータ登録に応じてデータ間の関係が随時変化していきますので、ニーズにあわせて柔軟に変化していくDBを利用者が育成することができます。

また、登録済みのコメントに対してコメントを追記していくことで、投稿型掲示板のようなコミュニケーション機能として使用することもできます。コメントなどの入力欄には日本語も入力できます。

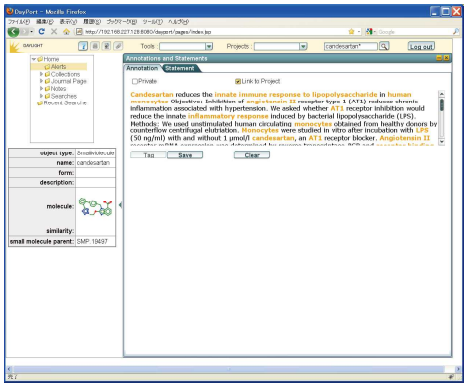

図4: アノテーションツール

■豊富なデータソース

DayPortのDBには、公共のデータソースを活用してあらかじめ1千万件ほどのデータが登録されています。データソースとしては、ChemBankやEMBL、GO（Gene Ontology）、SBO（Systems Biology Ontology）、MeSHなどを利用しています。生物学データや毒性データ、化合物や核酸・タンパク質情報、実験データやパスウェイデータなど、公共文献から抽出したセマンティック情報を収録しています。

社内に蓄積されている独自のインハウスデータも活用することで、独自の知識体系を構築できます。

Daylight社では高速な構造検索に代表される独自技術を基軸に、最新のWeb技術などを取り込みながら柔軟性の高いツールやシステムの開発を進めています。