バイオデータベースの XML 標準形式の設計と XML データ検索システムの開発

１．2002 年度の具体的な研究計画

（１）バイオデータベースの XML 標準形式の設計

医療や創薬の分野では、ゲノム医療・ゲノム創薬の要求の高まりから、ゲノムから始まり、細胞、器官、臓器を経て個体のレベルに至る幅広い範囲のデータが必要とされている。これに伴い、データベースの数も急速に増えつつあり、現在約 500 個近くのデータベースが存在すると言われている。これだけの数のデータベースを検索する場合、従来のようにユーザが個別にデータベースの場所を探索してそこからデータを検索し、得られたデータを手作業で対応付けていくのは労力の点で困難であり、ソフトウェアによる自動的なデータベース探索とデータの検索および検索結果の統合が必要になる。しかし、現状では例え、データベースの場所がわかり、必要なデータが検索できたとしても、それらのデータの表現形式がデータベースごとにまちまちで、相互のデータを対応付けるのは専門家の介在なしには困難である。そこで、個別のデータベースの書式に依存しない、標準的なデータ形式を設計し、各データベースの検索結果をこの標準形式に変換することにより、データ相互の関連性を明確にできるシステムの開発を目指した。

（２）検索用配列分類データベースシステム

近年配列データベースは急激に成長している。エントリが増大することは良いが一方で、機能予測や構造予測を行うためにデータベース検索を行う上では弊害が出てきている。例えば、HIV のプロテアーゼを問い合わせ配列として検索を実行すると(NCBI の PSI-BLAST サーバで nr に対して検索すると)、上位 1000 個のアラインメントが全て HIV のプロテアーゼ（つまり自身との）アラインメントに占められる。データベース検索方法の感度の問題ももちろん重要だが、生のデータベースを用いた構造/機能予測のためのデータベース検索はいずれ破綻をきたすことをこの結果は示唆している。

そこで、本システム開発では、①配列の類似性や、モチーフなどの様々な切り口で配列データ（１次データ）をクラスタリングすることで、冗長性を排除し、②各クラスタのメンバを表示する際に２次情報（アノテーションや構造分類など）を付加することで、高機能化を図ることを目的として、各クラスタや２次情報をグリッド上に分散配置したものをシームレスに連携できるようなシステムを構築することを最終目標とする。

これを踏まえて 2002 年度は、①クラスタリングの基準を“配列の類似性”とし、配列をいくつかのドメインに分けて近縁なグループに分類した上で、検索結果はその分類群の代表のみを出力することで、冗長性を排除し、②さらに構造（PDB）や構造分類（SCOP）などの２次情報を付加して高機能化を図ったシステムを構築し、グリッド上への展開を考慮した予備システム上にプロトタイプを実装することを目指した。

（３）薬物代謝情報 XML データの検索・表示システムの開発

創薬や医療分野において、目的とする化合物関連情報（レセプター情報、薬物代謝情報、毒性情報、副作用情報）をいち早く収集することが非常に重要である。

しかし、これら化合物関連情報は、それぞれの情報毎にデータベースが分散されており、一度に複数の情報を取り出すことができない。また、それぞれの情報は、標準化されておらず、文章形式になっているため、目的の情報を得るには、非常に時間と労力とノウハウが必要とされる。

そこで、バイオグリッドプロジェクトにおいて、化合物関連情報の標準化を推進していき、これらの情報を横断的に結び連携を図ることで、創薬、医療分野における連携統合のデータベースを開発し、容易ですばやく目的の化合物関連情報を得ることができるシステムをバイオグリッド上で展開することが、本プロジェクトの最終目標である。

2002 年度は、化合物の関連した情報の中で重要な情報の１つである薬物代謝情報の薬物代謝系ＸＭＬデータの検索・表示システムの開発を目指した。具体的には、薬物代謝情報データＸＭＬプロトタイプシステムの設計・開発を行い、検索性能などについて評価した。

（４）ペプチド関連情報データベースの XML 問合せシステム

現在、多くのデータベースサービスは通常 CGI ベースで提供されている。これは、個々のサービスを独自に構築していくうえでは適当な方法であるが、複数のデータベースを横断的に検索する目的には適していない。複数のデータベースを有機的にリンクして検索できるようにサービスを構築しているサイトも数多く見られるが、ユーザーが独自に検索対象を選択したり、横断的検索のための共通項目を設定することはできない。

本プロジェクトにおいてはデータの種類や形式の異なる多数のバイオ情報データベースを相互に連携させるため、XML をベースとするデータ標準形式に基づいて相互に連携してデータベースを検索するシステムの開発が目標となっている。しかしながら、このような新たな検索システムに対応させるために、すでに稼働している検索システム自体を変更するのは、労力の点から考えて困難である。また、インハウス・データベースシステムとして開発・販売されている検索システムを対応させることは不可能である。また、今後、多くの学術・研究サイトで提供されてくると考えられる新規のデータベースを統合していくためのシステムの開発は必須であるが、個々のデータベースサービスに個別に対応するのではなく中間コンポネントの追加で対応していくためのベースとなるシステムの開発を行う必要がある。

上記のような目的を達成するために、各データベースサイトですでに公開されている検索システムを大きく改造することなくウェブサービス対応をするためのシステムとして開発を行った。具体的には、蛋白質研究奨励会で現在稼働しているデータベースシステムを例に、これとユーザーの間にデータベース固有の形式と XML 標準形式との相互変換のレイヤーを開発する。標準形式をもとに問合せを行う際に、データベース固有の形式への変換ならびに問合せを発行し、結果を再び XML 標準形式に変換して問合わせ元に返すレイヤーシステムの開発を目指した。

（５）ネットワーク上での XML データ検索システム

データグリッドを実現するために必須となる要素技術の一つとして、複数の組織から公開されている XML 形式のデータを、ネットワークを介して相互利用できる環境の構築を目指した。初年度の具体的な目標として、バイオ分野において重要な位置を占めるタンパク

質データを公開しながらも、それぞれスキーマが異なるために実際の相互利用における運用性向上が求められている SWISS-PROT、PIR、PDB の各データベースを対象として、標準形式の XML データを利用したシームレスな検索システムを提供することを掲げた。

２．2002 年度の進捗状況と研究成果

（１）バイオデータベースの XML 標準形式の設計（**付録１参照**）

タンパク質を中心としたデータベースの標準形式の設計を行った。実際に SWISS-PROT, PIR, PDB という既存のタンパク質データベースをこの標準形式に変換したデータを利用することにより各々のデータベースの異種性を意識する必要がなくなった。さらに、これらのデータベースの標準形式を、問合せ時にネットワーク上で動的に連携させて結果を閲覧できる検索ツールを開発した。標準形式データベースと本ツールを利用することにより、複数のデータベースが仮想的に統合された連携データベース環境を実際に構築できることが実証できた。

（２）検索用配列分類データベースシステムの開発（**付録２参照**）

【相同性に基づいた配列の分類プログラムの開発】

配列分類のアルゴリズムとして、配列検索ツールである BLAST を利用して配列の相同性の高い領域をクラスタリングする方法を考案した。具体的には、BLAST 検索の結果出力される HSP を次々とつなぎ合わせていくことにより配列の中の領域をクラスターに分類する。このアルゴリズムを用いた配列分類プログラムを作製した。またこのプログラムを用いて、蛋白質の代表的なデータベースである SWISS-PROT のエントリの配列情報に基づくクラスタリングを行った。

【データベースシステムの開発】

今年度は世界の主要な蛋白質データベースについて、データ内容、スキーマ等の調査を行い、本データベースに２次情報を付加するデータベースとして、SWISS-PROT、PIR、PDB、SCOP を選定した。

配列分類情報、蛋白質の２次情報を保持するデータベースを RDBMS に実装し、本データベースに対して Web ブラウザからアクセスするための Web システムを予備システム上に構築した。本データベースシステムを通して SWISS-PROTt に対して BLAST 検索を行うことで、本データベースがその結果を解析し、ヒットしたクラスターを表示する。各クラスターとその構成メンバーについて、クエリーとヒットした領域と、それに関する２次情報の一覧を表示させることが可能である。

（３）薬物代謝情報 XML データの検索・表示システムの開発（**付録３参照**）

(3-1) 薬物代謝情報データ XML プロトタイプシステムの設計

①　対象データ及びデータベースの選定

薬物代謝情報のＸＭＬスキーマ設計及びシステムのコンテンツの項目は、薬物代謝情報として、すでにデータ項目が整備されているレンディック教授（ザグレブ大学）が開発し

た P450 代謝データベースをベースに設計した。

P450 薬物代謝情報として、化合物名、薬物代謝名、反応様式、反応タイプ、文献から構成される。また、サンプルデータとして、約３，７００件の薬物代謝反応データを登録、XML データベース化した。

② XML データの格納方法の検討

データの格納方法では、当初、XML データベースを用いての検索、RDB を用いての検索、全文検索技術を応用して３通りの直接検索手法の検討を行った。この中で、XML データベース(Xindice)を用いての検索、RDB(PostgreSQL)を用いて、２通りの検索手法で実装を試みた。

③ 検索方法の検討

検索方法に関して薬物代謝情報のみならず、薬物代謝酵素とタンパク質関連データベース(SWISS-PROT, PIR, PDB)の相互リンクも含めて検討を行った。その結果、薬物代謝情報の CYPname とタンパク質関連情報の Gene name とで項目名が一致していることを確認し、この 2 項目間で相互リンクを張る事により両データベース間の連携が可能になることを確認した。

(3-2) 薬物代謝情報データ XML プロトタイプシステムの開発

① XML コンバータの開発

本評価用プロトタイプシステムに格納するため、非 XML(タブ区切り)形式のデータを XML 形式に変換するためのコンバータ、およびまた非 XML(タブ区切り)形式のデータを RDB に格納するためのコンバータを開発した。

② 評価用検索機能の開発

評価用検索機能として、以下の項目を開発した。

・薬物、薬物酵素、反応タイプによるキーワード検索機能
・アウトプットの XML,HTML 選択機能
・薬物代謝情報から関連するタンパク質情報の検索および HTML,XML による出力機能

③ データベースの検索性能

当初 XML データベースの使用を考え実装を試みたが実用に耐えうる速度が得られなかった。原因としてはデータベース自体が Java で実装されている点などがあげられる。XML データベースに関しては今後の改良に期待している。

RDB を用いての実装では実用上問題無い性能が得られている。

④ データの有意義性

プロトタイプシステムにより薬物代謝情報とタンパク質関連情報の連携を確認できた。データベース間の連携を行う事で、各データベースに分散されている情報を取得できる様になり、データベースの利便性がより増すことが考えられ、今後レセプター情報、薬物代謝情報、毒性情報、副作用情報等の化合物情報、さらにはターゲットタンパクと化合物とのリンクを張る事で、データの利便性をより高めて行く様考えている。

（４）ペプチド関連情報データベースの XML 問合せシステム（**付録４参照**）

- データベースサービス統合化のための基礎となる中間レイヤーの開発を行った。これを構成することでデータベース管理システム・検索サービスシステムはそのままでウ

ェブサービス化が可能となる。また、中間レイヤーの設定ファイルを編集することで、インハウス・データベースシステムや他の公開されているウェブサービスもシームレスに統合し利用することが可能となり、さらに、中間レイヤーに変換機構を組み込むことで本プロジェクトにおいて開発される標準 XML に対応することができるとともに種々の検索エンジンを統合することもできる。

- 既存のデータベースサービスの例として、財団法人蛋白質研究奨励会の文献データベース(PRF/LITDB)とアミノ酸配列データベース(PRF/SEQDB)の検索サービスを行った。
- データ出力形式として従来の形式と標準化 XML 形式の両方に対応した。

（5）ネットワーク上での XML データ検索システム（**付録５参照**）

ネットワーク上での XML 検索システムの実装を完了した。具体的には以下の機能を提供している。

・ 標準形式 XML データの要素とそれに含まれる文字列からなる問合せ条件の指定によるキーワード検索、及び問合せ条件の追加による絞り込み検索機能
・ 一般の Web ブラウザから利用できる簡便なユーザインターフェイスの提供
・ SWISS-PROT、PIR、PDB のそれぞれの XML データのデータベースシステムへの一括追加機能
・ データベースサービスへの SOAP によるアクセス
・ 問合せ条件を満たすデータエントリの件数取得サービス
・ 問合せ条件を満たすデータエントリ取得サービス

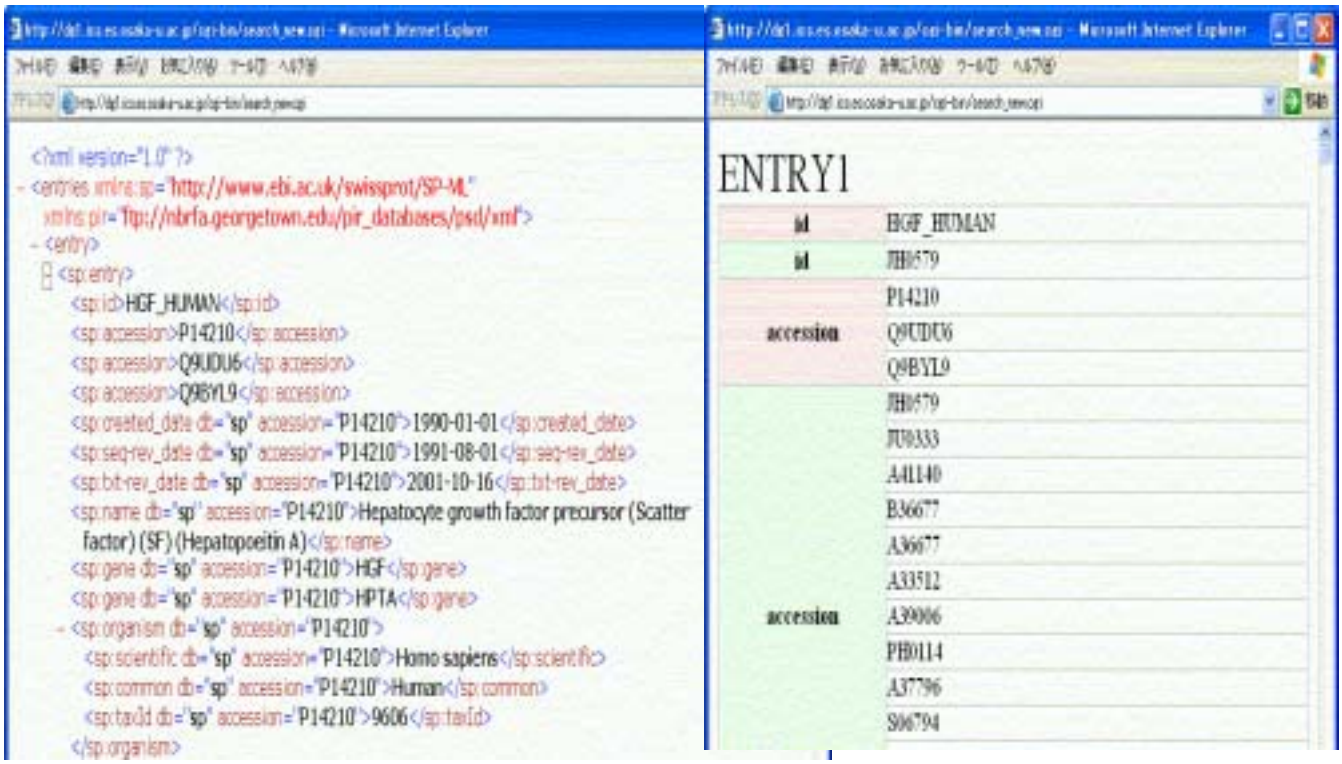
本システムを利用することで、SWISS-PROT、PIR、PDB の各 XML データを格納したデータベースに対して、その位置やスキーマの違いを意識せずにタンパク質データを検索することができる。

付録１　XML 標準形式

1.　標準形式への変換表

	Common Format	SWISS-PROT	PIR	PDB
エントリ	entry	entry	ProteinEntry	PDBj
エントリID	entry/id	entry/@name	ProteinEntry/@id	PDBj/@entry_ID
タンパク質名	entry/name	entry/proteinName	ProteinEntry/protein/name	PDBj/main/entity/entity_item/description
遺伝子名	entry/gene	entry/genes/gene/@name	ProteinEntry/genetics/gene/uid	PDBj/main/entity/entity_item/src_gen/gene
生物種(学名)	entry/organism/scientific	entry/organisms/organism/name	ProteinEntry/organism/formal	PDBj/main/entity/entity_item/src_gen/scientific_name
生物種(慣用名)	entry/organism/common	entry/organisms/organism/name	ProteinEntry/organism/common	PDBj/main/entity/entity_item/src_gen/common_name
文献	entry/reference	entry/references/reference	ProteinEntry/reference/refinfo	PDBj/main/citation/citation_item
機能	entry/function	entry/comments	--	--
EC#	entry/EC_num	entry/proteinName	--	PDBj/main/entity/entity_item/EC
キーワード	entry/keyword	entry/keywords/keyword	ProteinEntry/keywords/keyword	PDBj/main/struct/keywords
タンパク質配列	entry/sequence	entry/sequence	ProteinEntry/sequence	PDBj/main/entity/entity_item/seq_one_letter_code
タンパク質部分構造	entry/feature	entry/features/feature	ProteinEntry/feature	PDBj/struct/site
部分構造の型	entry/feature/type	entry/features/feature/@key	ProteinEntry/feature/feature-type	PDBj/struct/site/@id
部分構造の記述	entry/feature/description	entry/features/feature/@description	ProteinEntry/feature/description	PDBj/struct/site/site_gen/details

２．XML 標準形式の例



XML 表示　　XSLT による HTML 表示

付録２　電子計算機プログラム作成「検索用配列分類データベースシステム」

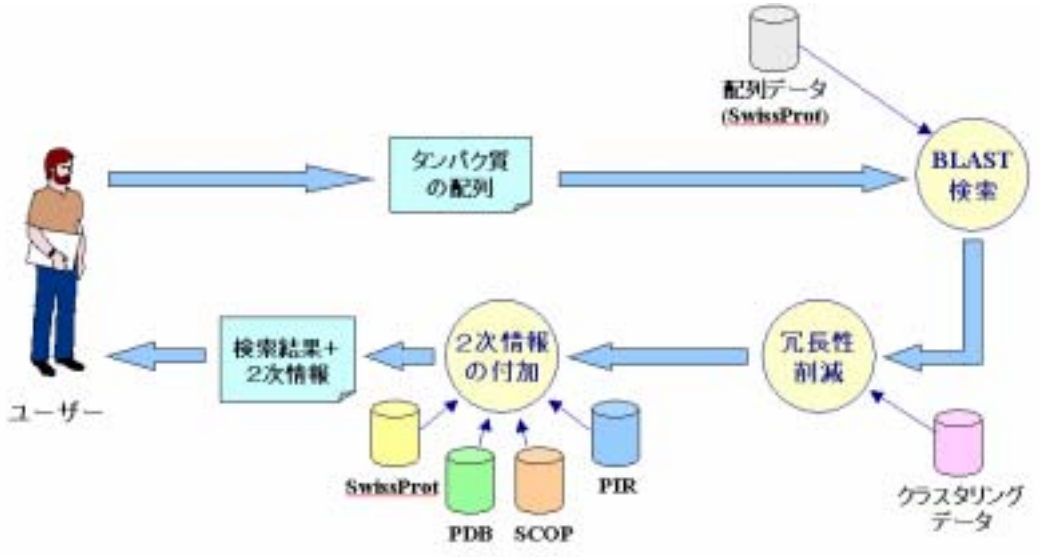
1. 検索用配列分類データベースシステムの概要

【HoCDB(Homology-based Clustering DataBase)とは？】

HoCDB は、配列データベースに対する BLAST 検索の結果を加工してユーザーに提供するためのシステムである。タンパク質や遺伝子の配列分類データベースは最も重要なバイオデータベースであるが、エントリの冗長性や、付加的な情報が少ないという２つの大きな問題があり、利便性が非常に低いの。HoCDB では、独自に作成したクラスタリングデータベースおよびタンパク質に関する各種公共データベースと、グリッド技術を連携することにより配列データベースの検索結果をよりよい形でユーザーである研究者に提供することができる。

【HoCDB での検索データの流れ】

検索データフローを下図に示す（最終的には各ＤＢをグリッド上に配置する）。



【クラスタリングデータ】

HoCDB のクラスタリングデータは、タンパク質の相同な領域をそれぞれのクラスターにまとめたものから構成される。配列データベースに対する生の検索結果では、相同なタンパク質が多数出力されてしまうという冗長性の問題があったが、クラスタリングデータを用いて冗長性を削減することが可能である。

【２次情報】

HoCDB では、各クラスターやメンバーに対して立体構造情報（PDB)、立体構造分類情報(SCOP)、タンパク質情報（SWISS-PROT、PIR）などの外部データベースからの２次情報を付加している。現在のバージョンでは２次情報ＤＢとの連携はまだグリッド化されていないが、来年度以降、２次情報ＤＢとグリッド技術を用いて連携することにより、常に最新の情報を取得することが可能となる。

2. 検索条件入力画面の説明

検索条件入力画面は、HoCDB のトップページであり、この画面から、データベースに対する BLAST 検索の条件を入力し、検索を実行することが可能である。

以下の各検索条件を入力し実行を開始する。

・期待値　…　検索の上限となる期待値を設定できる

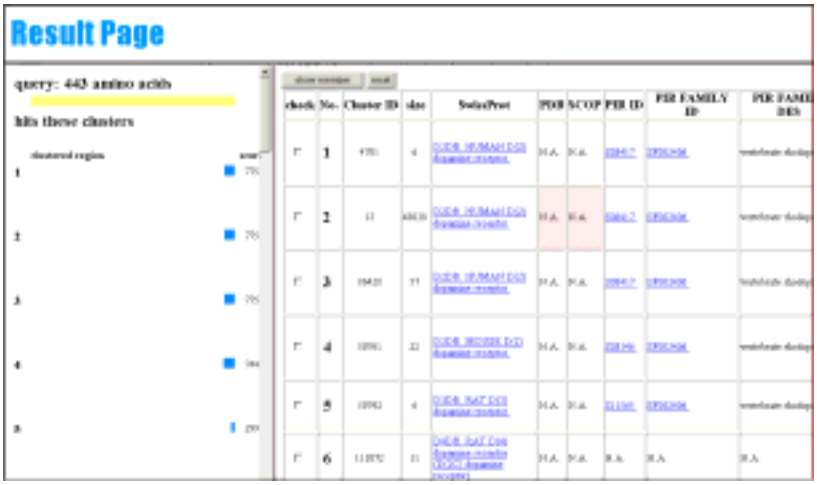
・アミノ酸置換行列の設定　…　検索配列と、データベース中の配列間の相同性スコアを算出するためのアミノ酸置換行列を設定することができる

・クラスターのサイズの設定　…　ある一定以下のメンバーのみからなるクラスターを検索の対象から除くことができる

・検索に用いる配列の設定　…　直接テキストエリアにペーストするか、またはファイルを選択することにより、検索配列を入力できる

3. 検索結果表示画面の説明

検索結果表示画面は、HoCDB に対する検索の結果を表示する画面である。



主に配列構造表示エリアと、配列情報表示エリアからなり、

・配列構造表示エリアには、最上部に検索配列長と検索配列の構造を表す黄色のバーが、その下に各クラスターの代表配列が検索配列のどこにヒットしたかを表す青色のバーが表示される。

・配列情報表示エリアには、クラスターに関する情報（クラスターＩＤ、クラスターのメンバー数）や代表配列の２次情報（PDB の立体構造情報等）が表示される。

4. メンバー情報表示画面

メンバー情報表示画面は、クラスターを構成するメンバーであるタンパク質の構造および情報を表示するものである。検索結果表示画面で、見たいクラスターを選択することでそのクラスターのメンバー情報表示画面を表示させることができる。

主にメンバー構造表示エリア（上図左）、メンバー情報表示エリア（上図右）からなり、

・メンバー構造情報エリアには、各タンパク質のどの領域がそのクラスターにメンバーとして登録されているのかが表示される。

左図の細いバーがタンパク質全長を表し、太いバーがクラスタリングされている領域を示している。

・メンバー情報表示エリアには、そのメンバーに関する２次情報が表示される。

付録３　電子計算機プログラム作成「薬物代謝情報 XML データの検索・表示システム」

1. システムの特徴

創薬や医療分野において、目的とする化合物の関連した情報（レセプター情報、薬物代謝情報、毒性情報、副作用情報）をいち早く収集することが非常に重要である。薬物代謝情報 XML データの検索･表示システムではこれらの情報の中で薬物代謝情報および薬物代謝酵素に関連するタンパク質の情報を取得する事ができる。今後レセプター情報、毒性情報、副作用情報と連携を行うことで、利用者にとってより使い易いシステムとなる。

2. システムの構成

薬物代謝情報 XML データの検索・表示システムの構成は以下の通りである。クライアントからの問い合わせが行われると、まず検索サーバが SOAP プロトコルを利用して各検索エンジンに対しリクエストを発行し検索結果を取得する。次に検索サーバ上で各検索エンジンから集められた検索結果に対し加工を行いクライアントに返す。なお、今回は評価のため１つのノード上ですべての機能を実現している。またタンパク質関連情報 XML としては SWISS-PROT のみを使用しており、PIR, PDB を含めた連携は来年度実施予定となっている。

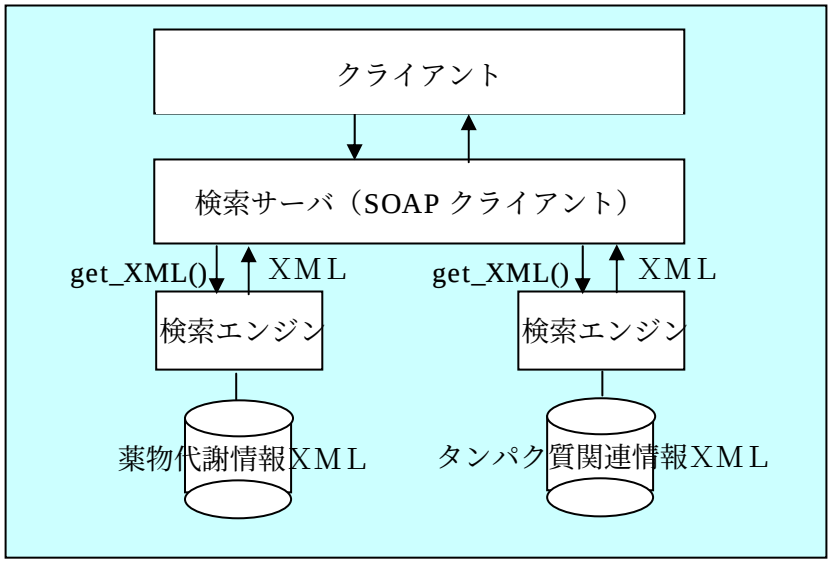


図 1. 薬物代謝情報 XML データの検索・表示システムの構成

3. 検索の流れ

薬物代謝情報 XML データの検索･表示システムにおける検索の流れは以下の通りである。まず最初に薬物代謝情報の検索を行う。薬物代謝情報 XML データの検索・表示システムでは薬物代謝酵素名、薬物名、反応タイプによるキーワード検索、またこれらのキーワードの AND,OR 検索が可能である。また出力タイプとして XML および HTML の指定を行う事ができる。

次に薬物代謝一覧画面(図 3)から HTML ボタンを選択すると薬物代謝情報と共に薬物代謝酵素名を元にタンパク質関連データベースの検索を行い、検索結果を併せて表示する(図

4)。XML ボタンが選択された場合はこれらの情報を XML で表示する(図 5)。

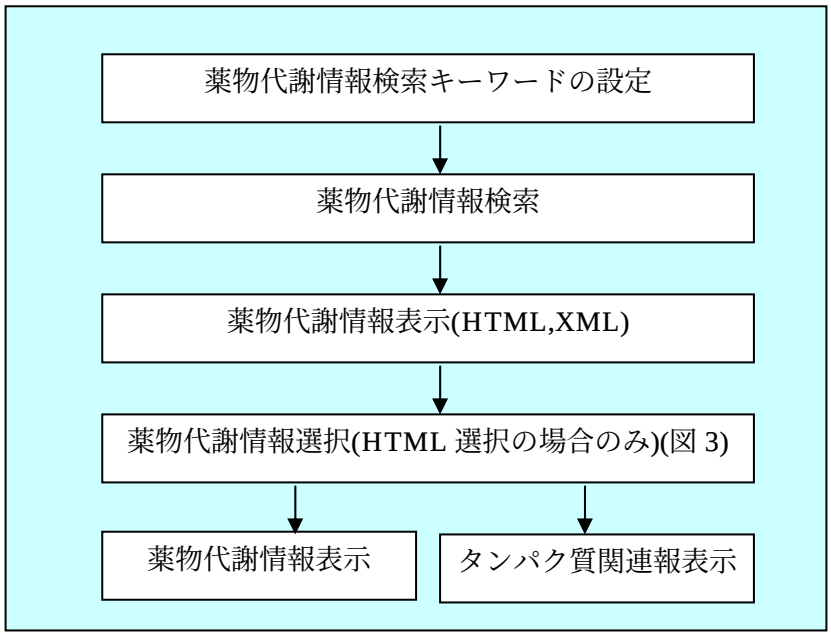


図 2. 薬物代謝情報 XML データの検索・表示システムの検索手順

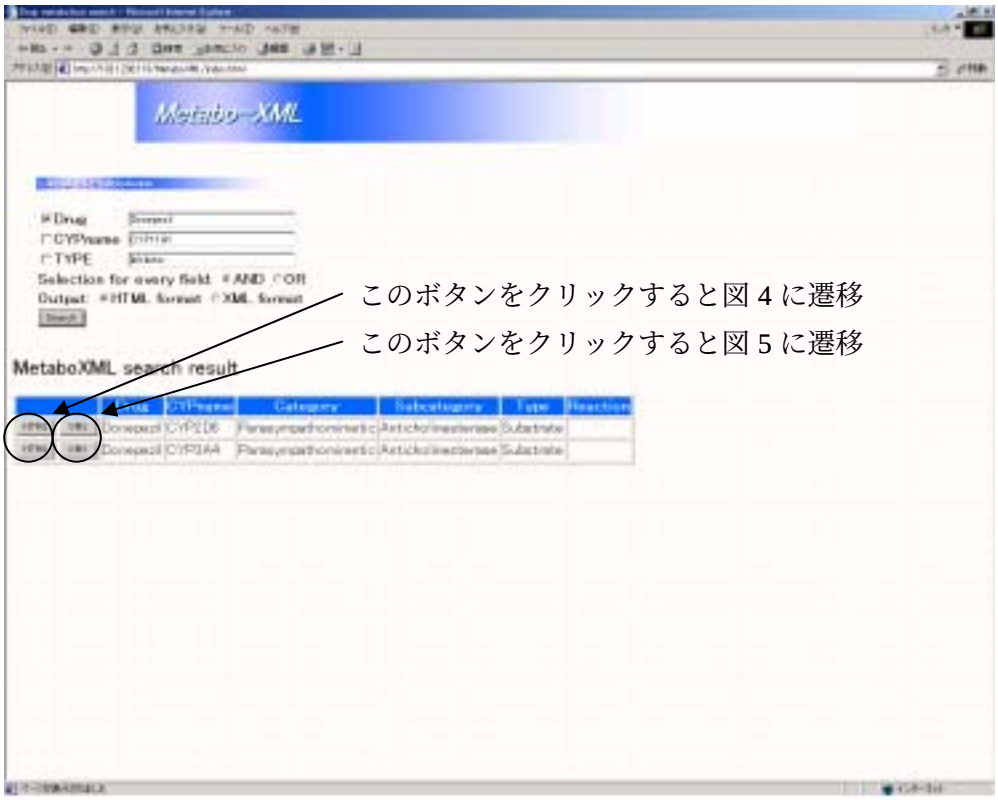


図 3. 薬物代謝情報の出力例(HTML)

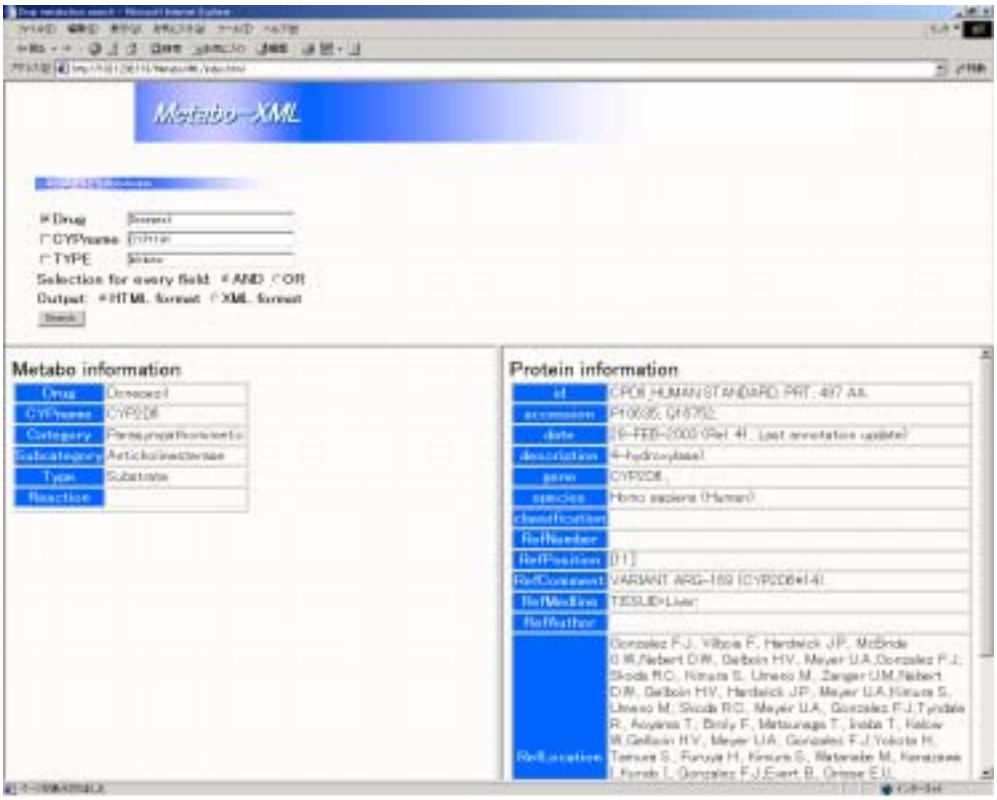


図 4. 薬物代謝情報およびタンパク質情報の出力例(HTML)

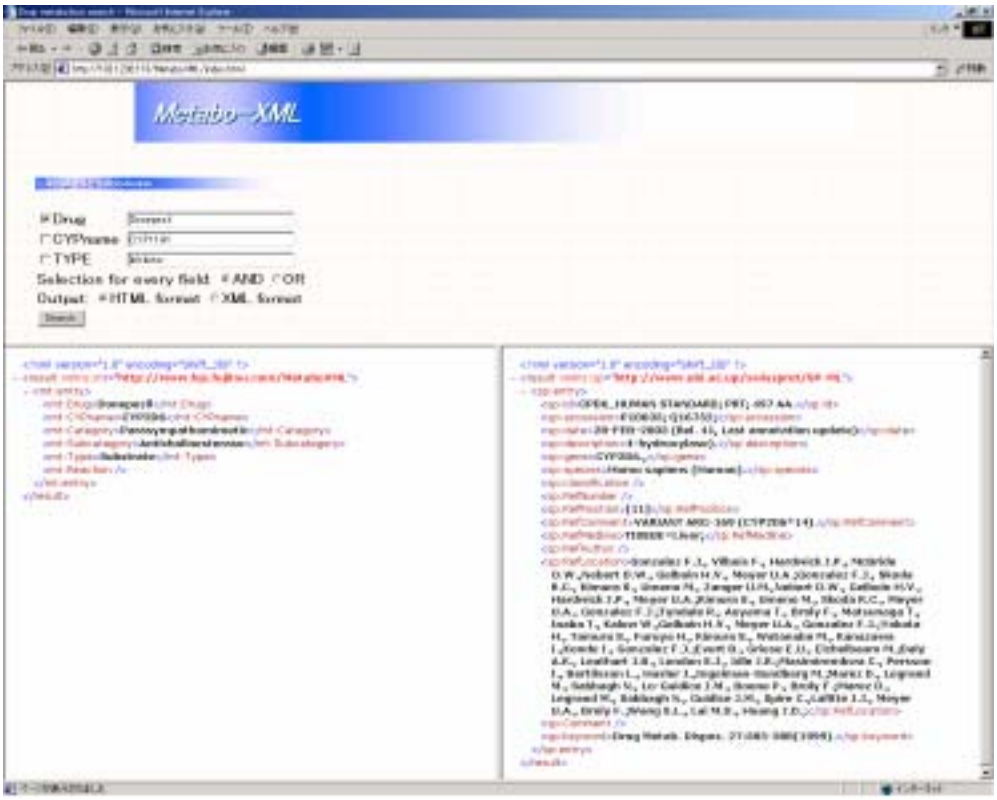


図 5. 薬物代謝情報およびタンパク質情報の出力例(XML)

付録４　電子計算機プログラム作成「ペプチド関連情報データベースの XML 問合せシステム」

1.　「ペプチド関連情報データベース」の XML 化とデータグリッド対応

現在、多くのデータベースサービスは通常 CGI ベースで提供されている。これは、個々のサービスを独自に構築していくうえでは適当な方法であるが、複数のデータベースを横断的に検索する目的には適していない。複数のデータベースを有機的にリンクして検索できるようにサービスを構築しているサイトも数多く見られるが、ユーザーが独自に検索対象を選択したり、横断的検索のための共通項目を設定することはできない。

データグリッド・プロジェクトにおいてはデータの種類や形式の異なる多数のバイオ情報データベースを相互に連携させるため、XML をベースとするデータ標準形式に基づいて相互に連携してデータベースを検索するシステムの開発が目標となっている。

しかしながら、このような新たな検索システムに対応させるために、すでに稼働している検索システム自体を変更するのは、労力の点から考えて困難である。また、インハウス・データベースシステムとして開発・販売されている検索システムを対応させることは不可能である。また、今後、多くの学術・研究サイトで提供されてくると考えられる新規のデータベースを統合していくためのシステムの開発は必須であるが、個々のデータベースサービスに個別に対応するのではなく中間コンポネントの追加で対応していくためのベースとなるシステムの開発を行う必要がある。

上記のような目的を達成するために、各データベースサイトですでに公開されている検索システムを大きく改造することなくウェブサービス対応をするためのシステムとして開発を行った。

例えば、財団法人蛋白質研究奨励会のデータベースサービスの仕組みは図１のようになっている。

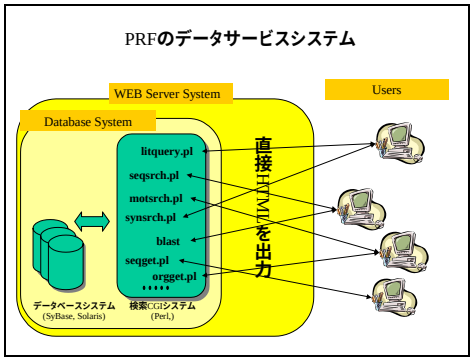


図１

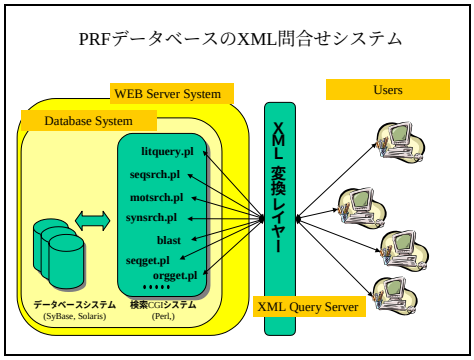


図２

図２のようにデータ転送の中間レイヤーを構成することでデータベース管理システム・検索サービスシステムはそのままでウェブサービス化が可能となる。また、中間レイヤーの設定ファイルを編集することで、インハウス・データベースシステムや他の公開されているウェブサービスもシームレスに統合し利用することが可能となり、さらに、中間レイ

ヤーに変換機構を組み込むことで本プロジェクトにおいて開発される標準 XML に対応することができるとともに種々の検索エンジンを統合することもできる。

2. 今年度の成果

- データベースサービス統合化のための基礎となるレイヤーシステムの開発を行った。
- 既存のデータベースサービスの例として、財団法人蛋白質研究奨励会の文献データベース(PRF/LITDB)とアミノ酸配列データベース(PRF/SEQDB)の検索サービスを行った。
- データ出力形式として従来の形式と標準化 XML 形式の両方に対応した。

3. システムの動作について

概要

ISAPI は IIS と同じメモリー空間（インプロセス）で動き、Filter と Extension の２タイプある。

まず、Filter タイプで URL の変換をする。

利用したいデータベースサービスサイトをあらかじめ登録しておく。財団法人蛋白質研究会のデータベースサービスを例として動作の概要を説明する。

例えば prf.or.jp -> prf として登録してあるとすると、
http://xxxx.com/prf/yyy.htm を http://xxxx.com/Scripts/Extension-name.dll に変換し、この DLL を介して http://prf.or.jp/yyy.htm から HTML を取得しブラウザに返送する。この結果、HTML が表示される。

この yyy.htm 内に<form method='POST' action='/Scripts/zzz.pl'>
の記載があって SUBMIT される場合、Filter 内で
<form method='POST' action='/Scripts/Extension-name.dll'>
に変換し、action='/Scripts/zzz.pl' という情報や、Form 内の変数情報などが ISAPI Extension である Extension-name.dll に渡される。

Extension-name.dll はこれらの情報を解析し、http://prf.or.jp/Scripts/zzz.pl に対して HTTP リクエストを出す。これはブラウザがする動作と同じであるから、Extention-name.dll は HTML の返信を受け取る。これをそのまま最初にリクエストを出したブラウザへ返さず、ここで、XML に変換しブラウザへ返送する。

こうすることにより登録された変換ルールに合わせて、CGI が返した HTML を XML に変換できる。

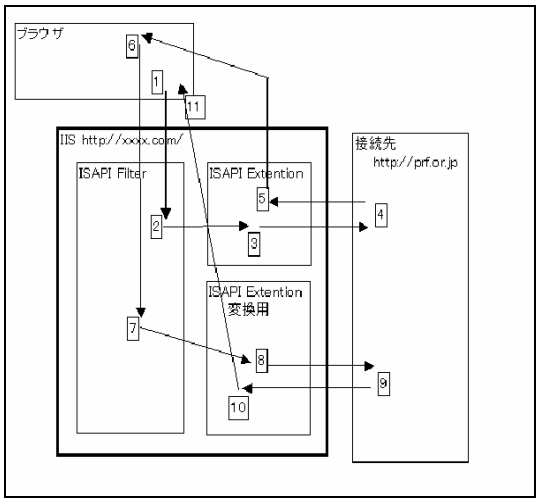
ウェブサービスに関しては、同様に例えば、http://prf.or.jp/Scripts/zzz.pl に対するリクエストを aaa.wsdl として登録しておき、http://xxxx.com/prf/aaa.wsdl を呼ぶと、内部で変換処理をして CGI を呼び、XML として返すことで実現可能となる。

以上の構成をとることにより、現状でデータベースサービスに使用しているデータベースシステム・検索システムにほとんど手を加えることなく WEB Service 化することができる。また、B2B システムに組み込むことで検索エージェントを開発することも容易になる。

また、本システムを IIS プラス ISAPI の組み合わせで実現することで、この仕組みを単なるプロキシとしてではなく、ローカルに案内ページを置き、そこに記載された URL か

らユーザーが簡単に、望むサイトにアクセスできるようになる。
例えば、テスト検索ページ
等を記載した HTML ファイルである。このことにより、自社のサイトを持たないユーザーも利用することができる。

XML 変換レイヤーの動作構成



1. ブラウザから HTML ファイルのリクエストを出す。
2. 登録してあるサーバーの URL に変換する。
3. Extension から外部のサーバーにアクセスする。
4. 外部のサーバーが HTML を返す。
5. そのまま、ブラウザに返信する。
6. ブラウザから、CGI にリクエストを出す。
7. 変換用の Extension にリクエストを送る。
8. 外部のサーバーへ CGI リクエストを出す。
9. HTML を返す。
10. HTML を XML に変換する。
11. ブラウザに返す。ここで XSLT 変換を行う。

4. 今後の目標

データグリッドにデータベースサービスシステムを容易に組み込むことができるようにするために以下の改良を行っていく。

(ア) 「異種・既存のデータベースサービス」をシームレスに統合するために、[OGSA-DAI]に対応する中間レイヤーの変換コンポネントの開発を行う。

(イ) 従来型の文字ベースのデータ以外に数値やグラフなどを扱う「ファクトデータベースに対応可能な中間レイヤーの開発を行う。

(ウ) 今後、新たに発表される「XML フォーマット」に対応するためのインターフェースの開発を行う。

付録５　電子計算機プログラム作成「ネットワーク上での XML データ検索システム」

1. ネットワーク上での XML 検索システムの目的

現在、バイオインフォマティクスに利用されるデータベースの種類は多岐に渡り、その形式も様々である。そのため、複数の異なるデータベースを利用して必要な情報を取り出すことが難しくなっている。我々が開発したネットワーク上での XML 検索システム（以下、本システム）は、グリッド基盤上に配置された XML 形式のタンパク質データベースを統合し、分散されたデータベースの所在やスキーマの違いをユーザに意識させることなく透過的に検索できるようにするシステムである。これは、プロジェクトの目標の一つであるグリッドデータベース基盤開発の基礎となることを目的とする。

2. システムの特長とソフトウェア構成

本システムでは、分散されたホスト上で提供するデータベース検索サービスを、SOAP を用いた Web サービスとして実現する。SOAP とは、インターネット上で公開されているサービスを HTTP と XML を用いて利用する仕組みであり、サービスのインターフェイスを XML で記述することで、プラットフォームに依存しないオープンなシステムを実現できる。次期グリッド基盤の標準と見込まれる OGSA は SOAP による Web サービスを基本アーキテクチャに採り入れており、本システムは SOAP を採用することで OGSA へのスムーズな移行をねらいとしている。

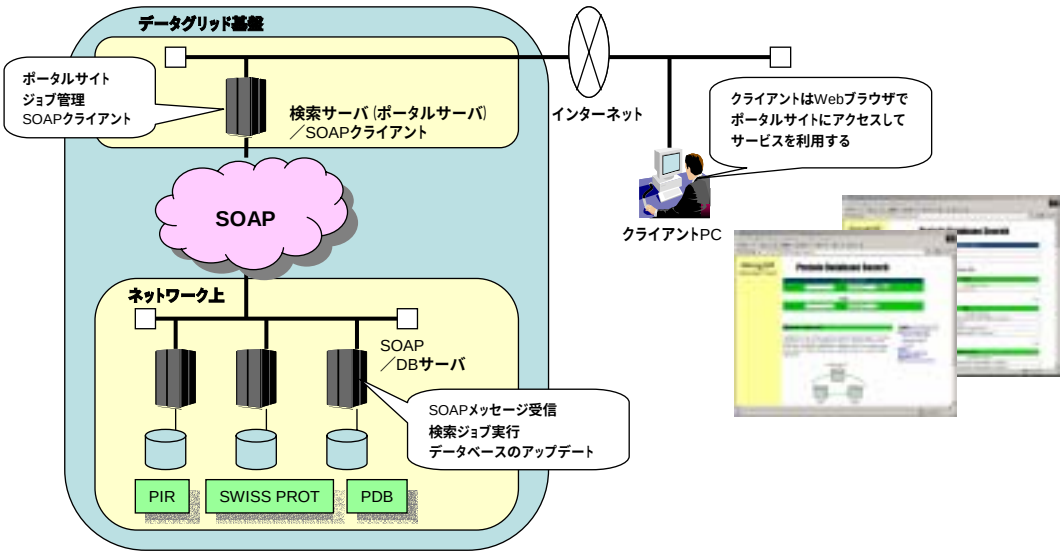


図１　システム概要

図１にシステムの構成を示す。本システムは検索サーバ、DB サーバ、XML データ格納の三つのサブシステムから構成される。検索サーバは本システムのポータルの役割を果たし、クライアントに対して XML 検索のユーザインターフェイスを提供する。DB サーバは XML データ検索サービスを提供する検索エンジンであり、SOAP によって外部からアクセスできるリモートプロシジャコールインターフェイスを提供する。XML データ格納は、日々

更新されているデータを含む XML ファイルの内容をデータベース管理システムに追加するバッチシステムである。ユーザがポータル（検索サーバ）に対して問合せを発行すると、ポータルはネットワーク上に分散配置されたホスト（ＤＢサーバ）上で提供される検索サービスを利用し、得られた検索結果を収集・加工してクライアントへ返送する。

2.1 検索サーバ

図２に検索サーバのソフトウェア構成を示す。検索サーバは、Web サーバ上で動作するサーブレットとして実装した。検索サーバはクライアントからアクセスされると、JSP で生成したユーザインターフェイス画面の HTML データを送る。ユーザは Web ブラウザからの簡便な操作で XML 検索機能を利用することができる。検索機能としては、検索条件を満たすエントリ一覧を取得する機能と、ID を指定したタンパク質データエントリを取得する機能を提供する。

XML 問合せ要求を受けると、検索サーバは問合せの内容に応じて複数の DB サーバ上で提供されている検索サービスを実行し、検索結果の XML データを得る。得られた XML データは XSLT を用いて HTML データに変換してクライアントに返す。

DB サーバ上で提供される検索サービスの情報は、DB サーバが稼動するホスト上で公開されている WSDL を参照する。検索サーバ上には、あらかじめ WSDL の URL を記述したファイルを保持しておき、ファイルに記述されている全ての URL を参照して WSDL を取得し、その内容を参照して検索サービスを利用する。

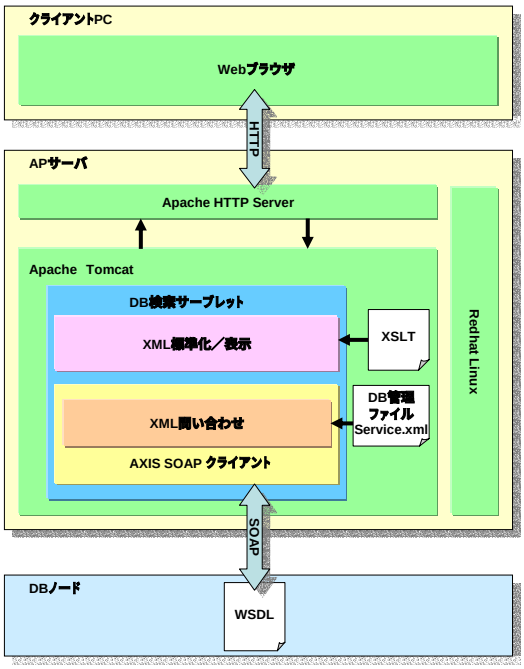


図２ 検索サーバの構成

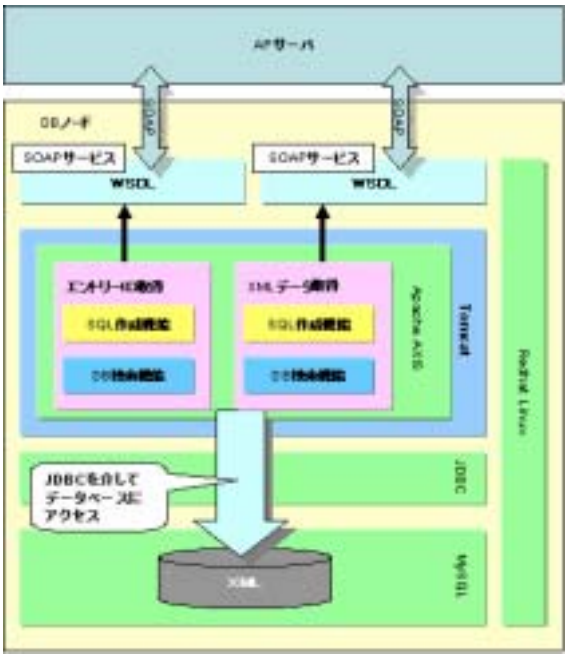


図３ ＤＢサーバの構成

2.2 DB サーバ

図３に DB サーバのソフトウェア構成を示す。本システムでは、XML データの取得に必要となる最も基本的な操作として、以下の三つの手続きを提供する。いずれも、SOAP を

介して外部からの呼び出しが可能な Web サービスとして実装している。

(1) エントリ件数取得

指定された検索条件を満たすエントリの件数を得る。

(2) エントリ一覧取得

指定された検索条件を満たすエントリの一覧（エントリ ID、タンパク質名）を得る。

(3) エントリデータ取得

指定された ID のエントリデータを XML テキストとして取得する。

「エントリ件数取得」と「エントリ一覧取得」は、標準 XML 形式のタグ名に基づく検索項目と検索キーワードの並びを引数とし、検索条件を満たすエントリの一覧を XML データとして検索サーバへ返す。「エントリデータ取得」は引数として指定したエントリ ID を持つタンパク質データを標準形式の XML データとして返す。DB サーバ上のサービスは Apache Axis 上で動作するサーブレットとして実装し、データベースエンジンには、パフォーマンスを考慮して RDBMS である MySQL を利用した。

2.3 XML データ登録

図４に XML データ登録サブシステムのソフトウェア構成を示す。XML データ登録は、SWISS-PROT、PIR、PDB のそれぞれに対応した標準形式の XML データフェイルを入力して、MySQL データベースにデータを格納するバッチ処理プログラムである。

各 DB ともに同一スキーマでテーブルを構築し、タンパク質１エントリの XML データを１レコードとして格納する。また、検索項目として利用する要素の内容を抽出してインデックステーブルを構成し、検索の高速化を図っている。

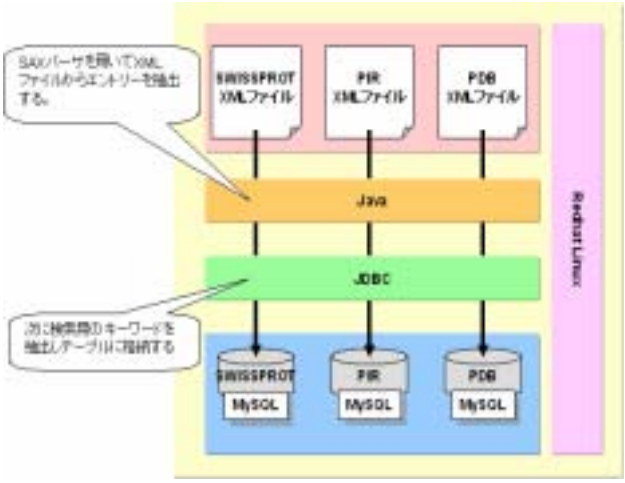


図４ XML データ登録の構成

3. 開発システムの目標達成度

本システムを実際に稼動させたところ、エントリ一覧の検索に数秒、エントリデータ取得に１秒未満と、実用的な時間内でネットワーク透過な検索機能が実現でき、データグリッドの基礎部分として利用できる見通しが得られ、当初の目標をほぼ達成できたと考えられる。