公共データベースの XML 変換システムの開発

１．2002 年度の具体的な研究計画

公共データベースは、それぞれデータの種類や形式が異なるため、それらを相互に連携して利用することが容易ではない。本年度は、いくつかの公共データベースにターゲットを絞り、これら定義が多様な公共データベースの内容をバイオグリッドプロジェクトで設計した XML ベースのデータ標準形式に基づいて半自動的に変換、表示するためのシステムの開発を実施した。具体的な開発項目は以下の通りである。

① 公共データベースから XML 標準形式への変換システムの設計
- 変換対象とする公共データベースの内容調査ならびに変換対象データベースの検討
- 公共データベースの XML インタフェースに関する調査と変換定義ファイルの設計
- 変換方法の方式提案

② XML 標準形式の表示システム設計
- 変換前の公共データベースと変換後の XML 標準形式の連携表示方法に関する検討
- 表示方法の検討

③ XML 標準形式への変換システムおよび XML 標準形式の表示システム開発
- 公共データベースから XML 標準形式への変換システムの開発
- エントリ ID 指定による XML 標準形式表示システムの開発

④ 評価作業
- XML 標準形式への変換精度
- XML 標準形式の表示性能

２．2002 年度の進捗状況と研究成果

① 公共データベースから XML 標準形式への変換システム設計
- 変換する公共データベースの内容調査と変換対象データベースの検討
  →対象の公共データベースを蛋白質データベースにし、PDB、PIR、SWISS-PROT を選定した。
- 公共データベースの XML フォーマット調査と変換定義ファイルの検討
  →選定したデータベースの XML フォーマットと XML 標準形式フォーマットから変換の際に対応表となる変換定義ファイルを設計した。
- 変換方法の検討
  →今後変換対象となるデータベース増加に対応できるよう、公共データベース毎に変換定義ファイルを用意し、ライブラリの追加により対応可能な方式を設計した。

② XML 標準形式の表示システム設計
- 公共データベースと XML 標準形式の連携表示方法の検討

→変換前後と変換エラーが視認性に優れた形で表示できるように、タグ移動や色変更等の HTML 機能を取り入れて設計した。

・表示方法の検討

→変換前後と変換エラー内容を統合して表示するために、標準的に用いられている Tomcat のサーブレット機能を採用した。

③ XML 標準形式への変換システム、変換結果の表示システム開発（**付録１参照**）

・公共データベースから XML 標準形式への変換システム開発

→各公共データベースごとに定義された変換定義ファイルを利用し、XML 標準形式へ変換するシステムの開発を実施した。

・エントリ ID 指定による XML 標準形式表示システム開発

→変換された XML 標準形式のエントリ ID を指定することにより、変換前後・エラー情報を統合して表示するシステムの開発を実施した。

④ 評価作業

・XML 標準形式への変換精度

→PDB、PIR、SWISS-PROT の各データベースから変換された XML 標準形式を任意に抽出し、変換精度を測定した。その結果、変換定義ファイルにて定義されている情報は全て変換対象として抽出できており、またタグ内の属性もタグへ展開するという要求を満足する抽出が実現された。

・XML 標準形式の表示性能

→XML 標準形式の表示性能の測定を実施した。

XML ツリー表示を Java にて実装しているためＸＭＬの行数が長くなれば、ツリーの伸縮時にレスポンスが悪くなるが現状での使用には耐えうる性能は有している。エントリ ID 表示の際、変換処理からのテキストログ情報を読み込んでいるため、変換件数が増加するに伴い速度が低下するので RDB 等によりデータベース化する必要がある。

付録１　電子計算機プログラム作成「公共データベースのＸＭＬ変換システム」

はじめに

ゲノム・蛋白質・化合物・医薬品等、複数のバイオデータベースの有機的連携を可能とするＸＭＬ化は、データベースの統合的利用において必須となる技術である。しかしながら、これらバイオデータベースは、種類やデータ量が多いだけでなく、厳密にスキーマが定められていないフラットファイルのデータベースやテキスト情報も多数含んでいる。また、書式や表記の揺らぎなどの曖昧性が存在するため、単純な置き換え作業ではＸＭＬ化は実現できない。

本ＸＭＬ変換システムは、これらの公共バイオデータベースで独自に提供されるインターフェイスを介して半自動的にＸＭＬ形式への変換と表示が出来るソフトウェアであり、XML 変換システムと XML 表示システムから構成されている。

XML 変換システムは、公共データベースで提供される各種データを XML 標準形式へ変換する目的で開発されており、データベースごとに定義された変換定義ファイルを利用したタグの置換機能、ならびに変換時のエラー情報をエラーレベルに対応させて保存する機能を持つ。XML 表示システムは、公共データベースのオリジナルデータと変換後の XML 標準形式データをエラー情報とともに連携表示する機能を持ち、変換時の不具合を定義ファイルにフィードバックする作業を効果的に支援する。

以下、これらのシステムの操作方法について説明する。

XML 変換システムの設定と起動方法

1.　設定方法

XML 変換システムはバイオデータベースを独自の XML 形式ファイルへ変換するシステムである。本処理を動作させるためには別途用意したバイオデータベースを所定のディレクトリへ設定する作業が必要になる。Ver1.0 においてはバイオデータベースの対象が PDBj-ML・PIR・SWISS-PROT となっている。

2.　起動方法

XML 変換システムはシェル画面からコマンド起動、またはシェルによるバッチ起動に対応している。通常は/home/XMLConvert/Bin に収録されている XMLConv.sh を実行することにより起動されるが、コマンド起動も環境変数とパラメータを指定することにより可能である。

・環境変数に関しては XMLConv.sh を参照。

・　パラメータについては下記を参照。

例）XMLConv GetDir=/home/Data/SP PutDir=/home/Data/SpOut DBTYPE=1

引数	名称	シンボル名	備考
第１引数	アプリケーション名	XMLConv	第１引数は固定
第２引数	DB 取得場所	GetDir=	ディレクトリ下のファイルを検索
第３引数	DB 格納場所	PutDir=	存在しない場合は作成
第４引数	バイオ DB 種類	DBTYPE=	１：PDB ２：SWISS-PROT ３：PIR

3. 起動後

処理が終了すると PutDir にて指定したディレクトリへ、バイオデータベースから変換された XML ファイルが作成される。

また/home/XMLConvert/Log へ変換処理を実施したログファイルが出力される。このログファイルは XML 表示処理にて使用するので、XML 表示処理を実施する場合は削除しないこと。

注意点：

XML 変換処理は起動時、XML ファイルの出力を指定したディレクトリに何らかのファイルが存在する場合、処理を起動しない仕様になっている。

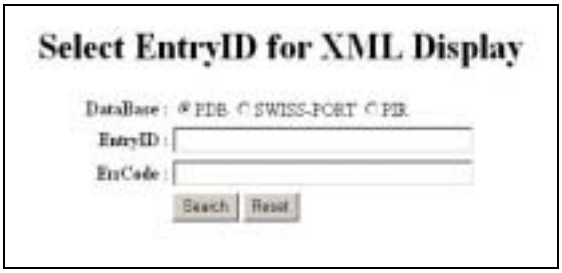
XML 表示システムの起動方法

1. 設定方法

XML 表示システムは Apatch の Web サーバ機能と Tomcat のサーブレット機能を利用したシステムである。これらの設定と起動が正しくないと XML 表示システムは正常に動作しないので注意する。Apatch と Tomcat の設定方法に関しては、各 URL を参照。

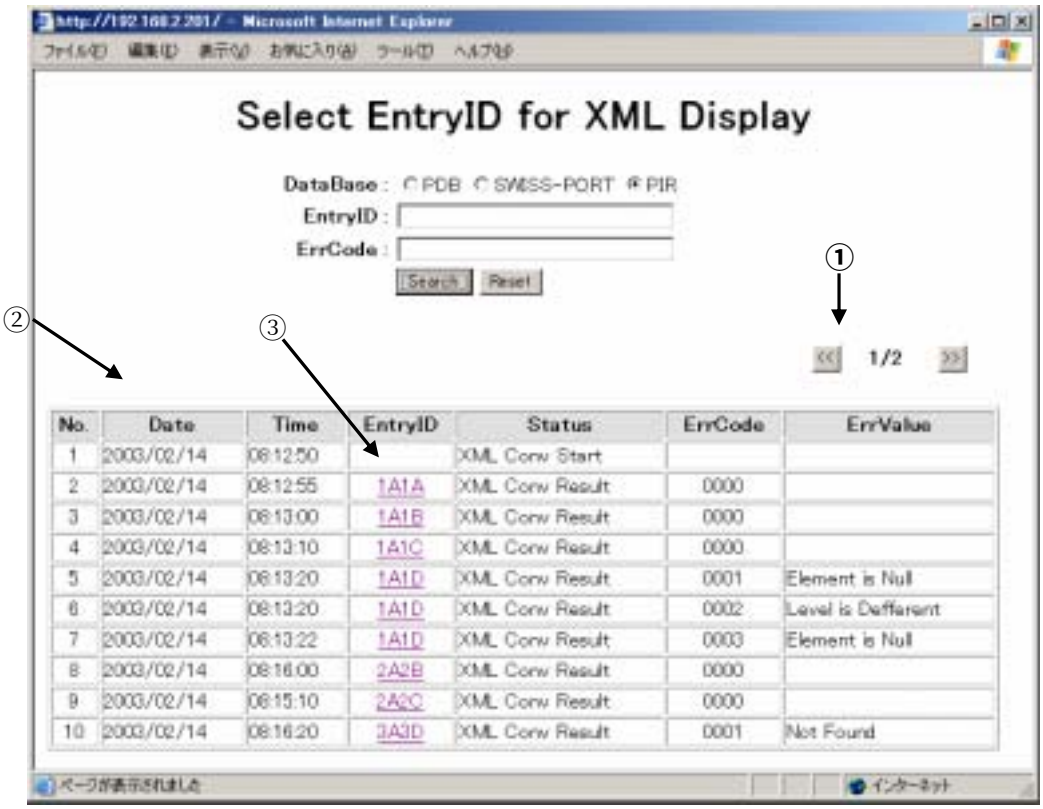
2. 起動方法

XML 表示システムを設定した PC のドメイン名または IP アドレスに xmldisp/index.html を付加した URL をブラウザに設定すると XML 表示処理が起動され下記画面が表示される。検索したい DB の EntryID または ErrCode を入力して検索を開始する。



検索結果画面

検索条件に基づいた EntryID 毎の内容が表示される。



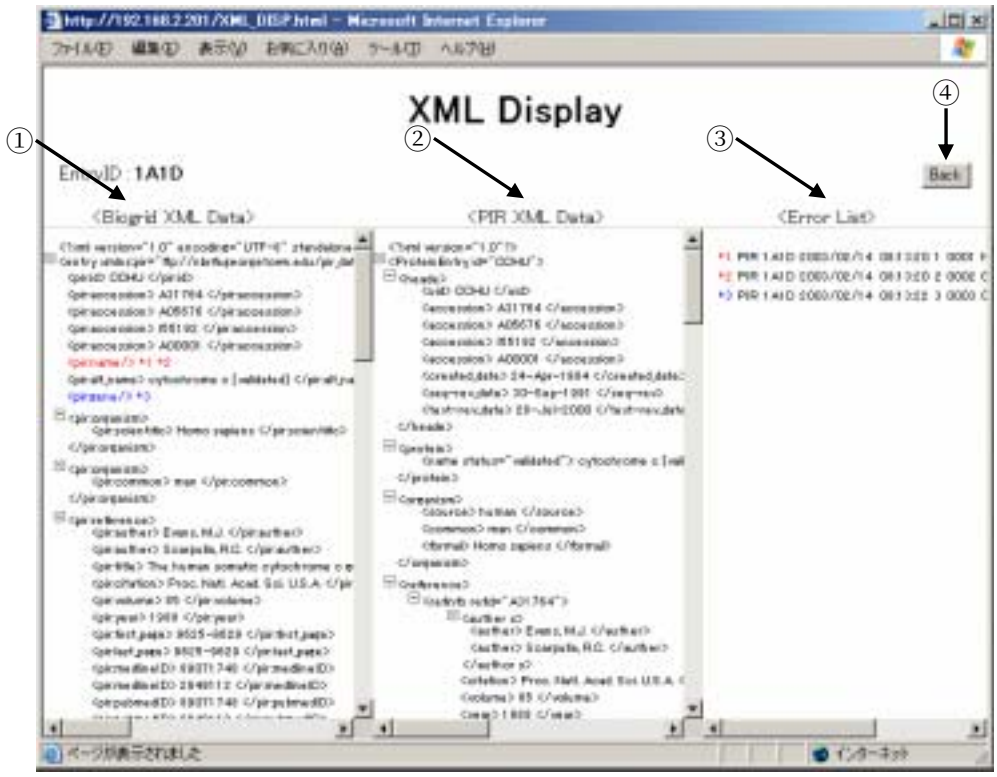
① 検索条件での検索結果ページ数と総ページ数を表示する。<<・>>ボタンを押下することにより前頁・次頁へと移動する。

② 検索条件での検索結果内容を表示する。検索結果は１画面最大１０行表示される。

③ 検索したい EntryID を選択することによりバイオデータベース XML 表示画面を表示する。

詳細表示画面

検索結果画面にて EntryID を選択することにより EntryID に対応するバイオデータベース・変換された XML ファイル・変換エラー内容が表示される。



① バイオデータベースから変換された XML ファイルを表示する。
② バイオデータベースの内容を表示する。
③ バイオデータベースからの変換時に発生したエラー情報を表示する。
④ Back ボタン押下により検索結果画面を表示する。

①～③の各フレームには拡張表示機能として、XML のタグを選択することにより他フレーム内に該当する表示箇所を、フレーム内先頭へジャンプすることができる。