ライフサイエンス分野における Web サービス

国立情報学研究所および日本ＩＢＭ東京基礎研究所
浦本直彦 (uramoto@jp.ibm.com)

1. はじめに

次世代の Web の基盤として登場した Web サービスは、現在、関連仕様の標準化やビジネスアプリケーションを中心とした普及が進んでいる。SCM、 ERP、 CRM といったビジネス分野における Web サービスの利点は、コンポーネント化したサービスインターフェイスを、簡単に結合し呼び出すことができることである。一方で、大規模なデータや知識を統合するための枠組みとしても Web サービスは使い勝手のよい基盤であり、ライフサイエンスなど、大量の情報が蓄積され、情報の統合が重要視されている分野においても、有望な手法として注目を集めている。本論文では、主に、情報統合の立場から、ライフサイエンス分野における Web サービスへの期待や適用例について述べる。

2. 情報統合基盤としての Web サービス

Web サービスは、Web 環境におけるコンポーネントおよびアプリケーションの統合のための基盤技術である。広義には、HTML、 Java Script、 JSP、 サーブレットなどの従来技術に基づくものも含むが、特に、Simple Object Accessing Protocol (SOAP)、 Web Services Description Language (WSDL)、Universal Description、Discovery and Integration of Web Services (UDDI) に代表される、XML に基づく標準技術を用いて構築されたものを指す。Web サービスでは、ネットワーク上のコンポーネントを、標準化された入出力を持つ「サービス」として定義し、それらを組みあわせることで、より複雑なサービスを構築することができる。

Web サービスの利点については、すでに様々な場所で言及されている。ここでは、以下の２点を挙げる。

1) サービスインターフェイスの統合。WSDL を用いて入出力の型を定義することで、ユーザは遠隔手続き呼び出し(RPC)や、XML メッセージを使って、簡単にサービスを利用することができる。また、これらのサービスを合成することで、ビジネスアプリにおけるワークフローを構築することが可能となる。

2) 情報・データの統合。分散された環境にある異種のデータを結合し、ひとつの仮想的なデータベースとして表現したり、異なる情報を関連つけたりすることができる。ユーザは、複数のデータベースに、それぞれのやり方でアクセスするのではなく、横断的に情報にアクセスし、統合された情報を入手することができる。

従来のWebサービスに関する議論では、主に(1)、つまり様々な環境にあるビジネスプロセスを容易に記述し、結合することの利点が強調されてきた。しかし、サービスの背後にある情報や知識を統合ための枠組みを与えることができるという(2)の利点もWebサービスの特徴のひとつであると考えられる。

そもそも、情報統合(データ統合)は、古くから人工知能やデータベースの分野で研究が進められてきた。Ullmanは、情報統合システムの共通アーキテクチャとして、ソース(source)、ラッパ(wrapper)、メディエータ(mediator)によるモデルを紹介している[1]。これに、関連する構成要素を加えたモデルを図1に示す。

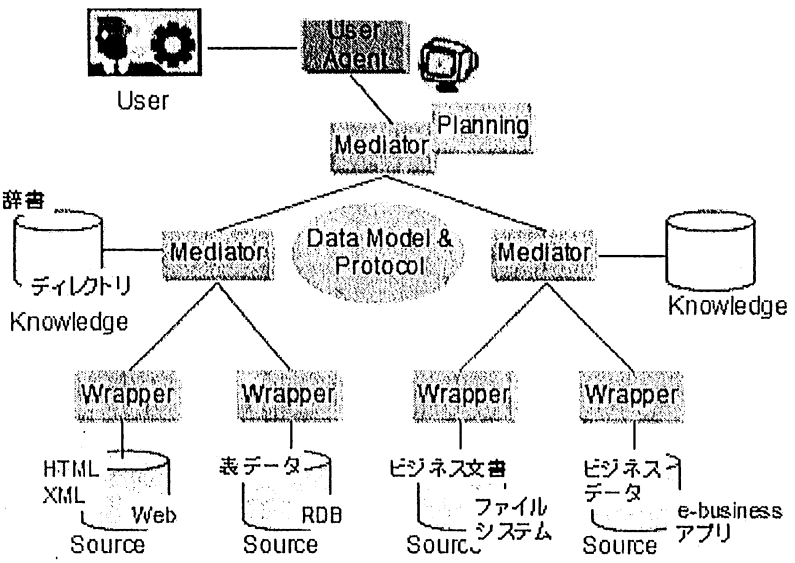

図1：古典的な情報統合のモデル

ソースは、それぞれ固有の(ローカルな)スキーマで記述された情報源を示す。XML文書におけるDocument Type Definition (DTD)やXMLスキーマ、関係データベースにおけるデータベース(表)スキーマ、アプリケーションが用いるデータの型などがローカルスキーマの例である。HTML文書のように意味的なローカルスキーマが定義されていない場合もある。

ラッパは、個々のソースを定義するローカルスキーマを共通のグローバルスキーマへ「翻訳」するために用いられる。グローバルスキーマは、各ソースの違いを吸収し、ソースの使用者に対して、共通のビューを提供するものである。Webサービスでは、サービス記述であるWSDL文書がグローバルスキーマに相当する。

メディエータは、グローバルスキーマへ翻訳されたソースを統合し、使用者からの要求に必要な情報を提供する。メディエータは、別のメディエータと連携することができる。また、メディエータを統合し複雑な処理を行うための戦略を受け持つプランニングエンジンが別に用意されることがある。

ラッパやメディエータが参照するグローバルスキーマに関する情報を供給するのが辞書、オントロジ、ディレクトリからなる知識源である。辞書はグローバルスキーマで用いられる単語あるいは概念の定義、オントロジは概念間の関係や制約(概念定義がオントロジに含まれる場合もある)、ディレクトリは、メディエータにアクセスするための情報(例えば所在情報やアクセス形式)を提供するために用いられる。単にサービスの型を統一するのではなく、サービスが取り扱う情報や知識をうまく統合しようと思うと、このような構成要素が基盤として必要になってくる。

3. ライフサイエンス分野からみたWebサービスの利点

ここで、ライフサイエンス分野がWebサービスに注目する理由について考えてみよう。

ライフサイエンス分野においては、初期の段階から文献情報や遺伝子、タンパク質に関する配列情報や数値情報が精力的に収集され、公開されてきた。例えば、米国生物工学情報センター(NCBI)では、ヒトゲノム計画の成果である遺伝子配列情報や、1960年代から1200万件に及ぶ医用文献抄録データベース(PubMed、 http://www。ncbi。nlm。nih。gov/entrez/)など、様々な形式の膨大な情報やツールを無料で公開している。NCBIを含め、多くの機関が有用な情報を整理して公共データベースとして公開しているのは、この分野の大きな特徴である。

このような大量でかつ様々な形式のデータを関連付けすることが非常に重要である。言い換えると、情報統合が本質的に必要不可欠であること[2][3][4][5]が、Webサービスを用いることの大きな要求となっている。例えば、ある遺伝子配列(A、T、 G、 Cの4種の塩基の並び)があるとしよう。この配列に対して、類似する遺伝子配列を検索する、あるいは、この遺伝子がコードしているタンパク質やそれに構造的あるいは機能的に類似するものを探す、そのタンパクについて言及されている文献を検索する、等々、ユーザは様々な情報源を渡り歩きながら、必要な情報を探し出す。データは文献、特許などのテキストデータだけでなく、配列データ、数値データ、画像データなど多岐にわたる。

さらにこれらの情報は、分散された環境に配置されていることが多い。よって情報源をいかに検索し、関連つけるかが大きな意味を持つ。個々のデータベースは、それぞれの研究機関で管理されていることが多いし、頻繁に更新されるので、人手による関連付けが非常に難しいからである。これらのサービスは従来HTMLベースで提供されていたが、情報の関連付けを行うには、XMLを使ってデータを構造化し、Webサービスを用いてサービスを統合するのは自然な流れである。データのXML化についても、前述のPubMedでは、全件の抄録データがすでにXML形式で閲覧可能である。また、京都大学化学研究所バイオインフォマティクスセンターが提供するDBGETデータベース(www。genome。ad。jp/dbget/dbget。links。html)は、複数のデータソースに対して、DNA、タンパク質、リガンドなどの仮想的なデータベースを定義し、単一の検索で、複数のデータベースを横断的に検索するための機能を提供している。個々のデータベースに対してはラッパに相当する機能雄用いて、複数データベースへの透過的なアクセスを実現している。

最後に、ライフサイエンス分野では、配列間の相同性（類似性に相当するもの）の計算や、タンパク質の構造および機能予測に大きな計算機パワーを必要とする点が挙げられる。これらの処理を

単一の計算機で実行するのは難しい。そこで、処理の並列化やグリッド化が必要となってくる。グリッド環境におけるミドルウエアとして、現在、Globus が脚光を浴びているが、最新のバージョンである Globus Toolkit 3。0 (http://www。globus。org/toolkit/)では、グリッド上のプログラムを、Web サービスとして定義し、ユーザが実際の計算機リソースがどのように使われるかを気にせずに用いることを可能にする。大阪大学下條らによる BioGrid グループ(www。biogrid。jp/)では、ライフサイエンス分野でよく使われるプログラムをグリッド上の Web サービスとして定義し、巨大な BioGrid 上で動かすことを提案している[5]。

4. Web サービスからみたライフサイエンスの利点

それでは、逆に Web サービスからみたこの分野の利点は何だろうか？

前述したように、情報統合では、セマンティクスを持ったデータを考慮することが必要である。これは、データのセマンティクスをできるだけ考慮することなくサービスの統合を実現しようとするＷｅｂサービスの理念とは相容れない場合がある。

しかし、前述の DBGET の仮想データベースでも示されているように、ライフサイエンス分野においては、遺伝子、タンパク質、アミノ酸、疾病、生物種、など、コミュニティの中で共通に使われる主要な概念が比較的少数である。うまく、共通のセマンティクスとデータモデルを設計することができれば、有効な統合システムを構築するすことができる(図２)。もちろん、同じタンパク質が異なる名前で呼ばれるなど、オントロジ（用語体系や同義語など）を整備していく必要があるが、PubMed における MeSH ターム(数万語レベルの概念体系)など、すでに利用可能な知識もいくつかある。

また、ビジネスアプリケーションにおいては、Web サービスのセキュリティが大きな問題となるが、ライフサイエンス分野においては、多くの公開データベースが存在するため、比較的統合システムを構築しやすいことも Web サービスが根付く土壌となりえる（もちろん、社内のデータと統合するような場合は、セキュリティの確保が重要であろう)。

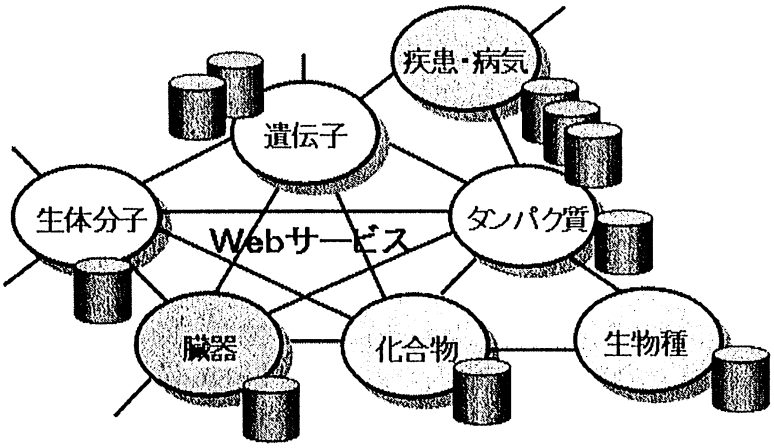

図2 ライフサイエンスにおける概念とサービス

5. 事例

ここまで、ライフサイエンス分野における Web サービスの重要性と適用のための枠組みについて述べた。本節では、実際の適用例をいくつかあげる。

人間の遺伝子配列を同定する国際ヒトゲノムプロジェクトのヨーロッパの拠点でもある European Bioinformatics Institute (EBI、 http://www。ebi。ac。uk/)では、遺伝子配列データを表現するためのXML構文である XEMBL を提案している。配列データは、CGI や Web サービスによって外部に公開されている。図3に XEMBL サービス記述(WSDL 文書)を示す。

ライフサイエンス分野でのソフトウエアの相互運用性を確保するために必要な技術を統一するために発足したコンソーシアムが、Interoperable Informatics Infrastructure Consortium (I3C、 www。i3c。org) である。現在、UDDI レジストリや Semantic Web[7][8]の利用などが検討されている。

日本遺伝学研究所生命情報・DDBJ 研究センターでは、ライフサイエンス分野でよく使われるツールを Web サービス化し公開している(http://www。xml。nig。ac。jp/index。html)。現在登録されているサービスを以下に示す。

- Blast
- ClustalW
- DDBJ
- ExClustalW
- Fasta
- GetEntry
- Gtop
- SRS
- TxSearch

これらのサービスの多くは、HTML ベースで公開されているが、Web サービスとして提供されることで、コンポーネントの一部として、より大きなシステム中に組み込むことができる。

IBMの先進ソフトウエア提供サイトである Alphaworks からは、Web Services for Life Sciences というツールが公開されている(http://www。alphaworks。ibm。com/tech/ws4LS)。提供されているのは、PubMed、 GenBank、 Blast、 Phylogenic Tree、 Clustal W に対する Web サービスである。

6. おわりに

本論文では、ライフサイエンス分野における Web サービスの重要性と適用事例について述べた。この分野における Web サービスの普及は始まったばかりであるが、これからの発展を期待したい。

```
<definitions name="XEMBL"   targetNamespace="http://www.ebi.ac.uk/XEMBL"
                            xmlns:tns=http://www.ebi.ac.uk/XEMBL" .... 一部略
<!-- 入力の型定義-->
<message name="getNucSeqRequest" xmlns:tns="http://www.ebi.ac.uk/XEMBL">
        <part name="format" type="xsd:string">
                <documentation>出力形式を指定するパラメータ </documentation></part>
        <part name="ids" type="xsd:string">
                <documentation>配列の識別子(アセッション番号) </documentation></part>
</message>
<!-- 出力の型定義 -->
<message name="getNucSeqResponse">
        <part name="result" type="xsd:string">
                <documentation>結果の XML メッセージ </documentation></part>
</message>
<!-- サービスオペレーションの定義 -->
<portType name="XEMBLPortType">
    <operation name="getNucSeq">
        <input message="tns:getNucSeqRequest" name="getNucSeqRequest"/>
        <output message="tns:getNucSeqResponse" name="getNucSeqResponse" />
    </operation>
</portType>
<!—結合情報の定義 -->
<binding name="XEMBLServiceBinding" type="tns:XEMBLPortType">
    <soap:binding style="rpc" transport="http://schemas.xmlsoap.org/soap/http" />
        <operation name="getNucSeq">
          <soap:operation soapAction="http://www.ebi.ac.uk/XEMBL#getNucSeq" />
        <input>
            <soap:body use="encoded" namespace="http://www.ebi.ac.uk/XEMBL"
                    encodingStyle="http://schemas.xmlsoap.org/soap/encoding/" /></input>
        <output>
            <soap:body use="encoded" namespace="http://www.ebi.ac.uk/XEMBL"
                    encodingStyle="http://schemas.xmlsoap.org/soap/encoding/" /></output>
        </operation>
</binding>
<!-- サービスポイントの定義 -->
<service name="XEMBLService">
    <port name="XEMBLPort" binding="tns:XEMBLServiceBinding">
        <soap:address location="http://www.ebi.ac.uk:80/cgi-bin/xembl/XEMBL-SOAP.pl" />
     </port>
 </service>
</definitions>
```

図 3:　XEMBL.のサービス記述 (WSDL 文書)

参考文献

[1] Ullman, J., "Information integration using logical views", Theoretical Computer Science, pp.189-210, Vol.239, No.2, 2000.

[2] Stein, L., INTEGRATING BIOLOGICAL DATABASES, Nature Reviews Genetics, pp. 337-345, Vol.4, May 2003.

[3] Davidson, S. B., et al, K2Kleisli and GUS: Experiments in Integrated Access to Genomic Data Sources, pp. 512-531, IBM Systems Journal, 40(2), 2001.

[4] Wheeler, D., et al, Database resources of the National Center for Biotechnology Information, Nucleic Acids Research, Vol.29, No.1, 2001.

[5] 上田ら, メタデータを用いたバイオ情報データベース連携検索手法の提案, 科学技術フォーラム(FIT), 2003.

[6] 浦本, Webにおける情報統合ーセマンティックWebとWebサービスー, 情報処理学会誌, Vol. 44, No. 7, 2003.

[7] 浦本, "Semantic Web - 機械のためのWeb -". 人工知能学会誌 16巻3号, 2001.

[8] Semantic Web, World Wide Web Consortium, http://www.w3org/2001/sw/