XMLの生物情報への応用

○宮崎　智
高橋　幸治
菅原　秀明

An application of XML for the microbial biological information

○Satoru MIYAZAKI
Kohji TAKAHASHI
Hideaki SUGAWARA

Abstract

Thanks to the genome projects and the research for the bio-diversity, we can find a great amount of biological information in Web servers. Thus it is very useful for us to set up a common mechanism to retrieve and exchange biological information worldwide. One of the most fundamental parts of the information sharing is to design the extensible format for the target data in diverse domains of biology. We propose a format by use of XML for the description of the microbial biological information such as bio-chemical characteristics, DNA sequences and their annotations.

1. はじめに

WWW による情報公開が進むつれて、単なるリンクによる関連付けを越えて、より質の高いデータ共有が望まれるようになってきている。データベースからの情報発信のしやすさ、他システムとのスムーズな情報交換を考えると、加工がしやすく、様々な付加情報を盛り込むことができると言う点では、XML による環境の整備は非常に魅力的であると言える。

ところで、生物系の情報は、多くの場合、階層的に管理することが可能であるが、データを表現するための項目は、実験技術や新しい知見の獲得とともに、頻繁に更新・削除が発生する。また、物理的に一つの研究所が総ての生物情報に関与することは到底無理であるから、分散環境での運用は避けられない。このような背景を考慮し、我々は、酵母の分類・同定用データを対象にしたXML形式を考案したので、ここに報告する。

2. 生物情報について

本研究は、生物情報の中でも、微生物の同定・分類データを対象としている。これらのデータに特徴的なことは、

1）学名は、いくつかの菌株をまとめるために便宜上つけられたラベルに過ぎず、名称の変更が頻繁におこる。
2）炭素化合物、窒素化合物の資化性や成長に必要なビタミンの情報など、いくつかの記号や比較的短い文字列で表現されるものと、遺伝子配列情報のようにテキストに近いもの、形態をあらわす画像データなど、データ種に富んでいる。
3）学名も含めて、実験手法の変遷とともに、管理上必要となる項目名の変更が頻繁におこる。
4）遺伝子配列情報以外は、特に、国際的基準フォーマットが制定されていないので、個々の機関が各々の仕様でインターネット上でデータ公開を行っている。

などが考えられる。このような背景で公開されたデータ群を共有して、統一的に扱うためには、それぞれの現状のフォーマットをなんらかの基準フォーマットに変更することが必要となる。フォーマットの定義を別に持つことができ、表現するための書式も付加情報として盛り込むことができるXML形式を導入することは、最少限の変更で、既存の異フォーマットデータを統一することができるであろう。

3. 酵母データの XML 形式によるフォーマット

2.節で述べたように、生物学的情報は、更新頻度が非常に高く、中でもデータ項目の変更も比較的的よく起こりうることが特徴的である。しかしながら項目を整理していくと、いくつかの項目間では階層構造が保たれることが多い。例えば、酵母の分類・同定用データの生化学試験の結果では、試験項目をいくつかの階層に分類することができこの分類の構

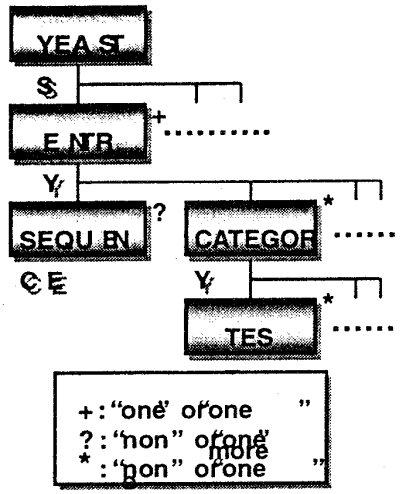

図 3-1. 酵母データの階層構造

```
<?xml encoding="US-ASCII"?>

<!ELEMENT YEASTS (ENTRY+)>
<!ELEMENT ENTRY (SEQUENCE?, CATEGORY*)>
<!ATTLIST ENTRY
        ID              CDATA   #REQUIRED
        NAME            CDATA   #REQUIRED
        COLONIES        CDATA   #IMPLIED
        GENUS           CDATA   #IMPLIED
        SPECIES         CDATA   #IMPLIED
        INPUT-DATE      CDATA   #IMPLIED
        MODIFY-DATE     CDATA   #IMPLIED
        TYPESTRAIN      (TRUE|FALSE)    'FALSE'
        STATUS          (PUBLIC|PRIVATE)'PUBLIC'
>
<!ELEMENT SEQUENCE EMPTY>
<!ATTLIST SEQUENCE
        ACCESSION       CDATA   #IMPLIED
        VALUE           CDATA   #IMPLIED
>
<!ELEMENT CATEGORY (TEST*)>
<!ATTLIST CATEGORY
        ID              CDATA   #REQUIRED
        NAME            CDATA   #REQUIRED
>
<!ELEMENT TEST EMPTY>
<!ATTLIST TEST
        ID              CDATA   #REQUIRED
        NAME            CDATA   #REQUIRED
        VALUE           CDATA   #IMPLIED
        DATE            CDATA   #IMPLIED
>
```

図 3-2. Document Type Definition

成は比較的安定的である。実際の試験項目は、これらの分類項目の下位にするデータ概念を考えればよい（図 3-1）。また、各試験項目の種類は、２００項目以上となる。XML 形式では、項目を１つのタグとして表現して導入されるデータ種をより厳密に検証することも可能であるが、処理速度および項目の削除を考慮して、項目は、タグの属性として導入することにした。図 3-2 は我々が考案した DTD である。このような構造にすることによって対応する項目の試験データが計測されていない場合は、XML 化されたデータの量を減らすことができる。この結果 JAVA1.2 で採用されている XML クラスを用いた場合、読み込みの速度を大幅に改善することができた。図 3-3 に XML 化された酵母データを示しておく。

```
<?xml version="1.0"?>
<!DOCTYPE YEASTS SYSTEM "YEASTS.dtd">
<YEASTS>
  <ENTRY ID="NB-78" NAME="NB-78" COLONIES=" "
         GENUS="Sporobolomyces" SPECIES="roseus" TYPESTRAIN="TRUE"
         INPUT-DATE="1999/04/08" MODIFY-DATE="" STATUS="PUBLIC">
    <SEQUENCE ACCESSION="AB006695"
         VALUE="gaaggctggg gaaagctcta gcctggcaga gtttcactcc ttttacatgct
                cacctttcag gatccccctt agcctgggc agcacagtcc tctactggga
                gaagcagtat cagtggagag">
    </SEQUENCE>
    <CATEGORY ID="1" NAME="Category-No1.">
      <TEST ID="1" NAME="C.Glucose" VALUE="+" DATE="1999/04/08"></TEST>
      <TEST ID="2" NAME="C.Galactose" VALUE="D" DATE="1999/04/08"></TEST>
      <TEST ID="3" NAME="C.L Sorbose" VALUE="D" DATE="1999/04/08"></TEST>
      <TEST ID="4" NAME="C.Sucrose" VALUE="+" DATE="1999/04/08"></TEST>
      <TEST ID="5" NAME="C.Maltose" VALUE="+" DATE="1999/04/08"></TEST>
      <TEST ID="6" NAME="C.Cellobiose" VALUE="+" DATE="1999/04/08"></TEST>
      <TEST ID="7" NAME="C.Trehalose" VALUE="+" DATE="1999/04/08"></TEST>
      <TEST ID="8" NAME="C.Lactose" VALUE="-" DATE="1999/04/08"></TEST>
      <TEST ID="9" NAME="C.Melibiose" VALUE="-" DATE="1999/04/08"></TEST>
      <TEST ID="10" NAME="C.Raffinose" VALUE="W" DATE="1999/04/08"></TEST>
      <TEST ID="11" NAME="C.Melezitose" VALUE="+" DATE="1999/04/08"></TEST>
      <TEST ID="13" NAME="C.Sol. starch" VALUE="+" DATE="1999/04/08"></TEST>
      <TEST ID="14" NAME="C.D Xylose" VALUE="D" DATE="1999/04/08"></TEST>
      <TEST ID="15" NAME="C.L Arabinose" VALUE="" DATE="1999/04/08"></TEST>
      <TEST ID="16" NAME="C.D Arabinose" VALUE="D" DATE="1999/04/08"></TEST>
      <TEST ID="17" NAME="C.D Ribose" VALUE="D" DATE="1999/04/08"></TEST>
      <TEST ID="18" NAME="C.L Rhamnose" VALUE="-" DATE="1999/04/08"></TEST>
      <TEST ID="20" NAME="C.Glycerol" VALUE="D" DATE="1999/04/08"></TEST>
</CATEGORY>
    <CATEGORY ID="2" NAME="Category-No2.">
      <TEST ID="1" NAME="Growth 25" VALUE="+" DATE="1999/04/08"></TEST>
      <TEST ID="2" NAME="Growth 30" VALUE="-" DATE="1999/04/08"></TEST>
      <TEST ID="3" NAME="Growth 35" VALUE="-" DATE="1999/04/08"></TEST>
    </CATEGORY>
  </ENTRY>
</YEASTS>
```

図 3-3. 酵母データの XML 形式による表現

4. 分類・同定ワークベンチ

XML などを用いて、交換用あるいは共有用にデータを表現する目的の１つは、微生物の同定や分類を実際にコンピュータ上で支援するシステムを開発し、使用することである。我々は、これまでに微生物データの管理システムと同定・分類機能を合わせ持ったワークベンチを開発してきている（図 **4-1**）。

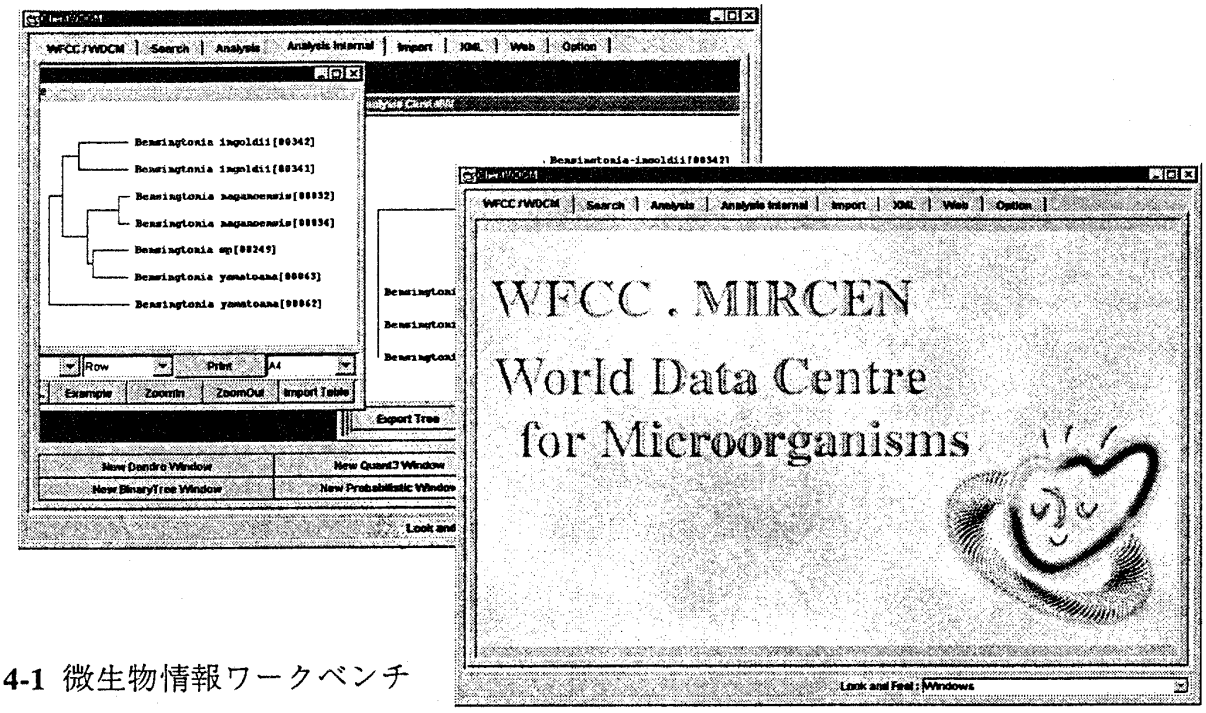

図 **4-1** 微生物情報ワークベンチ

先に述べたように、生物情報の大部分は安定的な階層関係を維持し得るので、ワークベンチの一部機能として、**OODB** によるデータベースシステムを構築した。

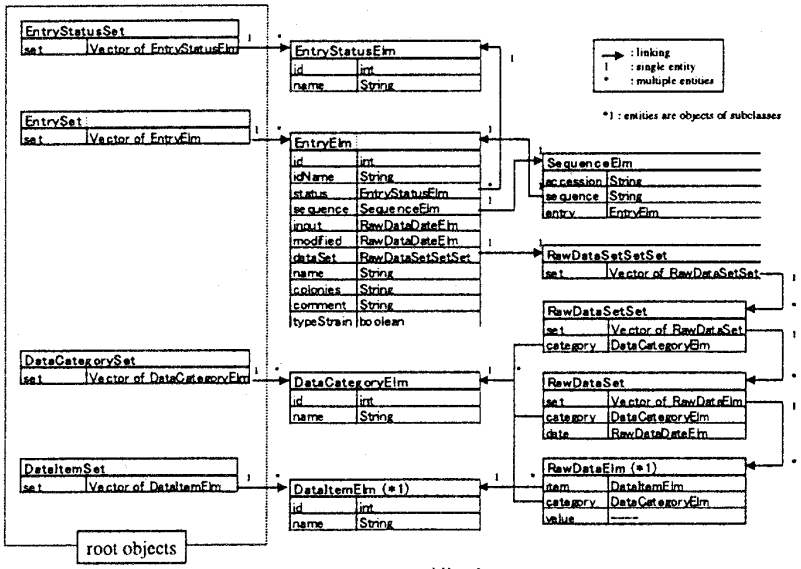

図 **4-2** 微生物ワークベンチのデータベース構造

一方、同定・分類用の解析部分では、ある菌株に関連するデータ群を一つのオブジェクトとして認識させ、必要に応じて一部を表示している。また、**3** 節で示した **XML** 化を通しても、データは菌株単位でオブジェクト化することができるので別途開発したワークベンチを比較的小さな変更で **XML** 形式のデータのインターフェースに改良することが可能とな

っている。こうして、生物情報を XML 形式化したものは、交換や HTML の代替としてだけでなく、既存の OODB の代替としても活用できることが示唆された。

5. まとめ

この報告に示した XML 化では、複数の菌株データを１つのファイルで XML 表現するものであった。しかしながら、レコード数が数万件以上となる場合には、１ファイルに総てのレコードを記述するのではなく、レコードごとに１ファイルを割り当て、全体の管理情報を別のファイルで記述するやり方がよいかもしれない。その場合は、全体の情報と個々のレコードのリンクをどのように実装するかあるいは実装できるかが問題となる。生物情報を記述するための基本的な XML 表現を検証することができたので、データの増加に伴うパフォーマンスについての知見を高めていく予定である。

参考文献

[1] Sugawara, H: Proposal of Information Biology. Iden 53, 39-43, 1999
[2] Sugawara, H. and Miyazaki, S.: An information system for data integration and polyphasic analysis of microbes. The Proceedings of the 99th General Meeting of the American Society for Microbiology, Chicago, 1999/5/30-6/3
[3] Miyazaki, S. and Sugawara, H. ,Reconstruction of large-scale phylogenetic trees: problems and solutions, Amino Acids Vol.17, No.1, pp119-120.
[4] Miyazaki, S and Sugawara, H., From linking to integration of biological databases, The International Joint Workshop for Studies on BIODIVERSITY (Species 2000, CODATA and Global Environment Tsukuba), Tsukuba, 1999/7/14-16
[5] Sugawara, H., Links between microbial resources and gene sequences, The proceedings of IXth International Congress of Bacteriology & Applied Microbiology,Sydney,1999/8/16-20
[6] Sugawara, H., Bio-resources are strategic information resources, the proceedings of Bio Resource Network Symposium, pp2-10, Tokyo, 1999/12/17
[7] Miyazaki, S. and Sugawara, H., Development of a prototype system for sharing and analysing data of microbial cultures, the proceedings of Bio Resource Network Symposium, pp36-45, Tokyo, 1999/12/17

宮崎　　智　国立遺伝学研究所（〒411-8540 静岡県三島市谷田 1111）
菅原　秀明　同上
高橋　幸治　日立ソフトウェアエンジニアリング（株）（〒231-0015 横浜市中区尾上町 5 丁目 79 番）

Satoru Miyazaki　(smiyazak@genes.nig.ac.jp)　National Institute of Genetics
Hideaki Sugawara　(hsugawar@genes.nig.ac.jp) National Institute of Genetics
Kohji Takahashi　(taka@lmi.hitachi-sk.co.jp) Hitachi Software Engineering Co., Ltd.