Veerasamy Ravichandran[1]
Joshua Lubell[2]
Gregory B. Vásquez[1]
Peter Lemkin[3]
Ram D. Sriram[2]
Gary L. Gilliland[1]

[1]Biotechnology Division
[2]Manufacturing Systems
 Integration Division,
 National Institute of Standards
 and Technology,
 Gaithersburg, MD, USA
[3]Laboratory of Computational
 and Experimental Biology,
 National Cancer Institute
 at Frederick,
 Frederick, MD, USA

Ongoing development of two-dimensional polyacrylamide gel electrophoresis data standards

We present an approach toward standardizing two-dimensional polyacrylamide gel electrophoresis (2-D PAGE) data in support of developing a globally relevant proteomics consensus in order to provide more efficient database querying and data comparisons through the establishment of the necessary definitions and interdisciplinary reference fields for both the 2-D PAGE community, particularly in the proteomics area, and the clinical and experimental biological research communities, in general. This article covers the need for unifying the 2-D PAGE data through a common data repository, and its usefulness in data standards and data interoperability.

1 Introduction

Two-dimensional polyacrylamide gel electrophoresis (2-D PAGE) has been an important tool for biological research for decades, but with the advent of proteomic research there has been a surge in both the number of users and the number of gels that are being run. The biological community marks the completion of the Human Genome Project's major goal in 2003: complete, high-accuracy sequencing of the human genome, which led to the discovery of genes coding for the production of tens of thousands of proteins [1, 2]. The current focus is on the gene products, specifically proteins, and the overall biological systems in which they act, creating the emerging fields of systems biology and proteomics. The free, widespread availability of a wide variety of data beyond human genome sequences – sequence variation data, model organism sequence data, organelle-specific data, expression data and proteomic data, to name a few – is starting to provide the means for scientists in all disciplines to better design and interpret their laboratory and clinical experiments, hopefully accelerating the pace of

biological discovery. Proteomics research is resulting in enormous amounts of data, orders of magnitude larger than the data generated in genomics studies, making the effective and efficient management of data essential. Having such a rich source of information is proving invaluable to scientists, whose findings should, in time, lead to improved strategies for the diagnosis, treatment and prevention of genetic diseases [3, 4].

However, it has become increasingly clear that simply generating the data is not enough; one must be able to extract from it meaningful information about the system being studied. Biological system data holds phenomenal promise for identifying the mechanistic basis of organismal development, metabolic processes, and disease, and it can confidently be predicted that bioinformatics research will have an additional impact on improving our understanding of such diverse areas as the regulation of gene expression, protein structure determination, comparative evolution, and drug discovery [5–9]. Of central importance to the optimal utilization of proteomics data, particularly 2-D data is the development of a unified infrastructure to facilitate collecting, storing, retrieving, and querying data, regardless of the technology used to generate it.

2 Need for data standards

Data standards are essential because they permit cooperative interchanges and querying between diverse, yet dissociated databases. The ability to interchange data in a seamless manner becomes critically important, and the economic benefits of data interoperability standards are immediate and obvious.

Correspondence: Dr. V. Ravichandran, Center for Advanced Research in Biotechnology of the University of Maryland Biotechnology Institute and the National Institute of Standards and Technology, 9600 Gudelsky Drive, Rockville, MD 20850, USA
E-mail: vravi@nist.gov
Fax: +301-738-6255

Abbreviations: HTML, hyper text markup language; **HUPO**, Human Proteome Organization; **PSI**, Protein Standards Initiative; **NIST**, National Institute of Standards and Technology; **SVG**, scalable vector graphics; **TWODML**, two-dimensional electrophoresis markup language; **XML**, extensible markup language; **XSL**, extensible style sheet language; **XSLT**, extensible style sheet language transformation

Proteomics and 2-DE

2.1 Evaluated/annotated data for data exchange

Lack of data standards is the Achilles heel of data interoperability. The lack of integration, implementation and use of standards are barriers to the delivery of optimal biological data. Even with the dramatic increase in the volume of proteomics data, innovation will be constrained unless technical advances are made in producing critically evaluated data and integrating data sources through data management, mining, and integration techniques. In the context of time-to-completion pressures and volumes of data, the research community needs certified models that can derive "best" recommended values from critically evaluated experimental data and validated and benchmarked predictive methods for any real or proposed measurement. These virtual measurement systems could generate data suitable for immediate use in commercial, scientific, and regulatory applications. Also needed are effective data management standards and techniques (*e.g.*, quality, traceability, or uncertainty estimates) for gathering, integrating, and maintaining information about data accessed from diverse sources [10]. Comparison of 2-D PAGE data between experiments and laboratories is essential, particularly as this data is utilized in clinical diagnostic settings. Accepted data standards for 2-D PAGE experiments would be a valuable asset for the study of diseases, especially when samples are rare and often unique. With respect to querying across biological databases, required information for a particular protein or family of proteins can be difficult to obtain from heterogeneous resources because of lack of data standards, leading to data compatibility problems. Thus, accurate and comprehensive automated information exchange between these resources is very limited.

Although public proteomics data resources are highly informative individually, the collection of available content would have more utility if provided in a standard and centralized context and indexed in a robust manner. Use of controlled vocabulary is already facilitating analysis of high-throughput data derived from DNA microarray experiments. The potential impact of improved interoperability derives from the fact that information and knowledge management systems have become fundamental tools in a broad range of commercial sectors and scientific fields [11–13]. The adoption of common standards and ontologies for the management and sharing of proteomics data is essential [14]. Even though, there exist many biological data-exchange formats (Table 1) [14], a well-documented and annotated data with easy exchangeable data format, such as an extensible markup language (XML) format, would help in data mining, annotation, storage, and distribution.

Table 1. Biological data-exchange formats

Format	Pros	Cons
Flat files	The simplicity of the flat file, which does lend itself to simple tools that are available to all, is in great part responsible for its popularity. Flat files, based on field/value, are commonly used to represent biological data.	Lacks referencing typed values vocabulary control constraints, among other issues. Often fields are ambiguous and their content is contextual.
Abstract syntax notation one (ASN.1)	ASN.1 files convey the description of their structure and offer some flexibility; the client side does not necessarily need to know in advance the structure of the data. Based on that single, common format, a number of human-readable formats and tools are produced, such as those used by Entrez, GenBank, and the BLAST databases.	ASN.1 software tools do not scale well for very large data sets, and there is no support for queries. Workers on ASN.1 standards recognized several years ago that there was a requirement (from users of ASN.1) to have an XML representation of the information structures defined by an ASN.1 specification.
Common object request broker architecture (CORBA)	CORBA provides platform-independent programming interfaces and models for portable distributed object-oriented computing applications. Its independence from programming languages, computing platforms and network protocols provides a solution for developing new applications for querying and distributing biological data, which can also be integrated into existing systems.	Despite the benefits of developing a CORBA environment, it's a heavy task that requires highly skilled computer scientists and expensive software. Further, CORBA messages are blocked by most firewalls, making CORBA an impractical option for Internet-accessible systems.

Table 1. Continued

Format	Pros	Cons
Java remote method invocation (RMI)	RMI enables the programmer to create distributed Java technology-based to Java technology-based applications, in which the methods of remote Java objects can be invoked from other Java virtual machines, possibly on different hosts.	Like CORBA, its scalability depends greatly on the network bandwidth.
Object-oriented database management system (ODBMS)	ODBMS extend the object programming language with transparently persistent data, concurrency control, data recovery, associative queries, and other database capabilities. Offers rich data model well suited to biology.	High learning curve to fully understand the technology:

2.2 XML for biological data exchange

XML was defined by the XML Working Group of the World Wide Web Consortium (W3C: http://www.w3.org/XML). XML is a markup language for documents containing structured information. Structured information contains both content (words, pictures, *etc.*) and some indication of what role that content plays (for example, content in a section heading has a different meaning from content in a footnote, which means something different than content in a figure caption or content in a database table). Almost all documents have some structure. A markup language is a mechanism to identify structures in a document. The XML specification defines a standard way to add markup to documents. XML is a simple, very flexible text format, playing an increasingly important role in the exchange of a wide variety of data on the Web and elsewhere [15]. Because an XML document so effectively structures and labels the information it contains, the web browser can find, extract, sort, filter, arrange, and manipulate that information in highly flexible ways. XML has been designed for ease of implementation and for interoperability with the World Wide Web.

The XML definition consists of only a bare-bones syntax. When an XML document is created, rather than use of a limited set of predefined elements, the data elements are created and subjectively assigned names as desired – hence the term extensible in XML. Therfore, XML can be used to describe virtually any type of document, and fits ideally with the requirements for the complex and diverse biological data integration. XML thus provides an ideal solution for handling the rapidly expanding quantity and complexity of information that needed to be put on the Web. Hence, a common language, like XML, should therefore offer power, scalability, adoptability, interoper-

ability, flexibility with different data types. In order to enhance the interoperability between diverse data, adoption of a universal data-exchange language, like XML, to exchange the annotated data would be useful. The number of applications currently being developed by biological communities that are based on, or make use of, XML documents is truly amazing. This is to facilitate the writing and exchange of scientific information by the adoption of a common language in XML (Table 2).

2.3 Markup language for electrophoresis data

Although there has been rapid progress in establishing standards for genomic sequence data as well as DNA microarray data, it is unlikely to be the case with proteomic data. Human Proteome Organization's (HUPO) Protein Standards Initiative (PSI) aims to facilitate standards for data representation in proteomics. Currently, the PSI is focusing on developing standards for two key areas of proteomics; mass spectrometry and protein-protein interaction data that will be XML-based. We propose here a common language for sharing electrophoresis experimental data, two-dimensional electrophoresis markup language (TWODML) that is based on the XML. Given the amount, complexity of data associated with a single set of electrophoresis experiment, we believe that XML is the most suitable method to describe electrophoresis data. The goal of the TWODML is to: (i) gather, annotate, and provide enough information that may be reported about an electrophoresis-based experiment in order to ensure the interoperability of the results and their reproducibility by others; (ii) help establishing public repositories and data-exchange format for electrophoresis-based experimental data; (iii) eliminate barriers to data exchange between the electrophoresis data, and permit

Table 2. Scientific markup languages

Markup language	Purpose	URL
Chemical markup language (CML)	Exchange of chemical information	http://www.xml-cml.org
Mathematical markup language (MathML)	Exchange of mathematical formula	http://www.w3.org/Math
Bioinformatic sequence markup language (BSML)	Exchange of DNA, RNA, protein sequences and their graphic properties	http://www.sbw-sbml.org/index.html
BIOpolymer markup language (BIOML)	Expression of complex annotation for protein and nucleotide sequence information.	http://www.bioml.com/BIOML
The taxonomical markup language	Exchange of taxonomic relationships between organisms	http://www.albany.edu/~gilmr/pubxml
Genome annotation markup elements (GAME)	Annotation of biosequence features	http://xml.coverpages.org/game.html
BlastXML	Model NCBI Blast output	http://doc.bioperl.org/releases/bioperl-1.2/Bio/SearchIO/blastxml.html
Ontology markup language/conceptual knowledge markup language (OML/CKML)	Representation of biological knowledge and specifically functional genomic relationships	http://smi-web.stanford.edu/projects/bio-ontology
Multiple sequence alignments markup language (MSAML)	Description of multiple sequence alignments (amino acids and nucleic acid sequences)	http://xml.coverpages.org/msaml.html
Systems biology markup language (SBML)	Representation and modeling of the information components in the system biology	http://www.cds.caltech.edu/erato/sbml/docs
Gene expression markup anguage (GEML)	Exchange of gene expression data, gene expression markup language	http://www.oasis-open.org/cover/geml.html
GeneX gene expression markup language (GeneXML)	Representation of the Gene Expression Databases datasets	http://xml.coverpages.org/geneXML.html
Microarray markup language (MAML)	Integration of microarray data	http://xml.coverpages.org/maml.html
Protein markup language (ProML)	Exchange of protein sequences, structures, and families-based data	http://www.bioinfo.de/isb/gcb01/talks/hanisch/main.html
RNA markup language (RNAML)	Exchange of RNA information	http://www-lbit.iro.umontreal.ca/rnaml/

the integration of data from heterogeneous sources; and (iv) leverage low-cost XML-based technologies such as XSLT (extensible style sheet language transformation) and SVG (scalable vector graphics).

The first step in data interoperability is to enforce standards for the electrophoresis data. A much bigger challenge is arriving at what descriptors constitute a 'required/minimum acceptable' set with respect to different types of parameters. The minimum information necessary from any 2-D PAGE experiment is that associated with the experimental details, in order to ensure firstly the reproducibility of the experiment, and secondly the interoperability of the results. An electrophoresis data standards committee can define a general-purpose set of TWODML elements, together with a document structure, to be used for a particular class of documents. By defining the vocabularies in a standard format (*e.g.*, the experimental sample source) the resulting uniformity may permit comparison of data between different systems (microarray data, macromolecular data, *etc.*). Hence, in the case of 2-D PAGE, the common data elements and data definitions for the required information were outlined.

The following data elements should be collected in association with their required data categories: sample source, detail about the protein, experimental detail, sample preparation, sample loading, sample separation condition, sample separation, experimental analysis, data analysis, and author information (Fig. 1)*. For example, the sample source information should contain the following tentative data items: the source record, which specifies the biological and/or chemical source of each molecule in the entry. Sources should be described by both their common and scientific names. Two types of sources will be grouped: the natural source (Fig. 2) and genetically modified (recombinant) source (Fig. 3). A part of recombinant source TWODML model data conforming to the schema fragment shown in Fig. 3 might appear as in Fig. 4. For example, the gene used in the recombinant sample source is described as below:

---

* Figures 1–4 were created using XMLSPY's schema designer. Rectangular boxes in these diagrams represent XML elements. The octagonal symbols are sequence compositors. Thus, Fig. 1 indicates that a TWODML root element shall contain a SAMPLE_SOURCE element, followed by an EXP_DESIGN element, and so on.
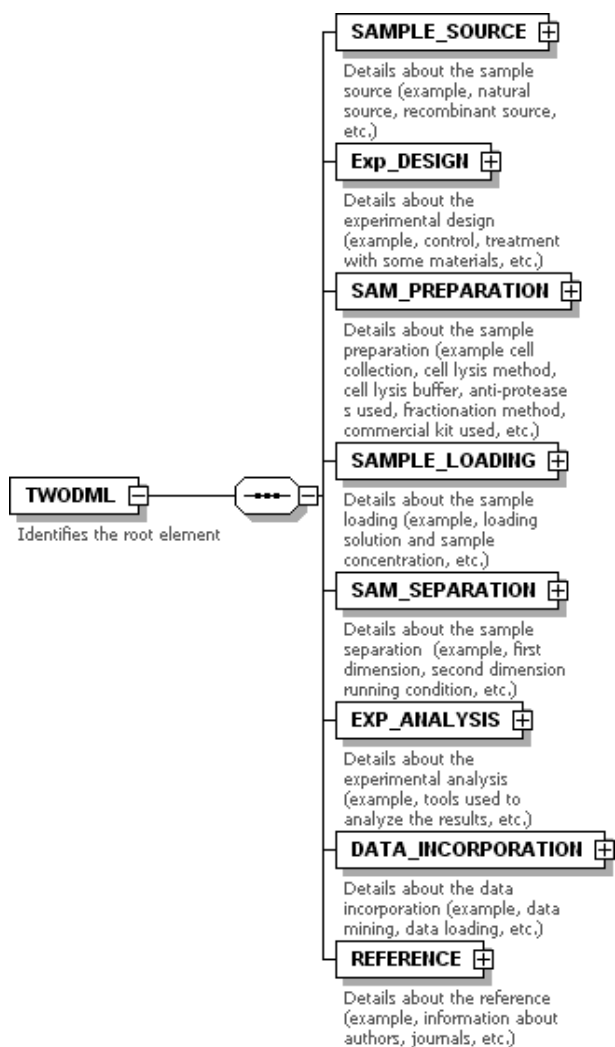
Figure 1. Schematic diagram for the TWODML.

```
<TWODML>
  <SAMPLE_SOURCE>
    <RECOMBINANT_SOURCE>
     <GENETIC_MATERIALS>
       <GENE>
          <NAME>SNAP-23</NAME>
          <PROTEIN_NAME>Snaptosome-asso-
          ciated Protein of 23 kDa
          </PROTEIN_NAME>
          <SOURCE>
           <ORGANISM_SCIENTIFIC>Homo
           sapiens
           </ORGANISM_SCIENTIFIC>
           <ORGANISM_COMMON>Human
           </ORGANISM_COMMON>
           <CELL_LINE>Raji – human B lympho-
           cyte (Burkitt's Lymphoma). ATCC num-
           ber: CCL-86
           </CELL_LINE>
          </SOURCE>
          <GENETIC_VARIANCE>Amino acid 23 is
          changed from Ser to Ala (Ser23Ala)
          </GENETIC_VARIANCE>
          <ORIGIN>
           <NAME>Dr. Roche</NAME>
           <ADDRESS>National Cancer Institute,
           NIH</ADDRESS>
           <CONTACT_INFO>pr17m@nih.gov</
           CONTACT_INFO>
          </ORIGIN>
          <MORE_INFO>GeneBank/EMBL Data
          Bank accession number: U55936
          </MORE_INFO>
       </GENE>
```

In this partial TWODML document, data elements are marked by enclosing sets of angle brackets. Standardized values for well-defined data elements are embedded within the elements. In addition to well-defined syntax described above, each TWODML document also carries information about: data relationships, data types, range restrictions, interdependencies, exclusivity, units, and methods. Free, open-source software is available to validate and parse the data files.

As in the natural source, data items in the genetically modified category record details of the source from which the sample was obtained. Associated data for this category include: the gene modified in the source material for the experiment, the genetic variation (transgenic, knockout), the system used to express the recombinant protein, and the specific cell line used as the expression system (name, vendor, genotype, and phenotype). Data items in

the natural source category will record details of the sample source. Associated data for this category will include: the common name of the organism and its scientific name, the source condition (normal, disease), any genetic variation, sex, age, organ, tissue, cell, organelle, secretion, and cell line information. The cell line and strain should be given for immortalized cells when they help to uniquely identify the biological entity studied.

The TWODML application is being defined by creating an XML schema which defines and names the elements and attributes that can be used in the document, the order in which the elements can appear, and constrains the values of elements and attributes. A variety of schema languages is available for XML. The oldest and simplest of these schema languages is the document type definition (DTD) [16]. Although DTDs enjoy strong software support, they are inadequate for representing strongly typed or context-
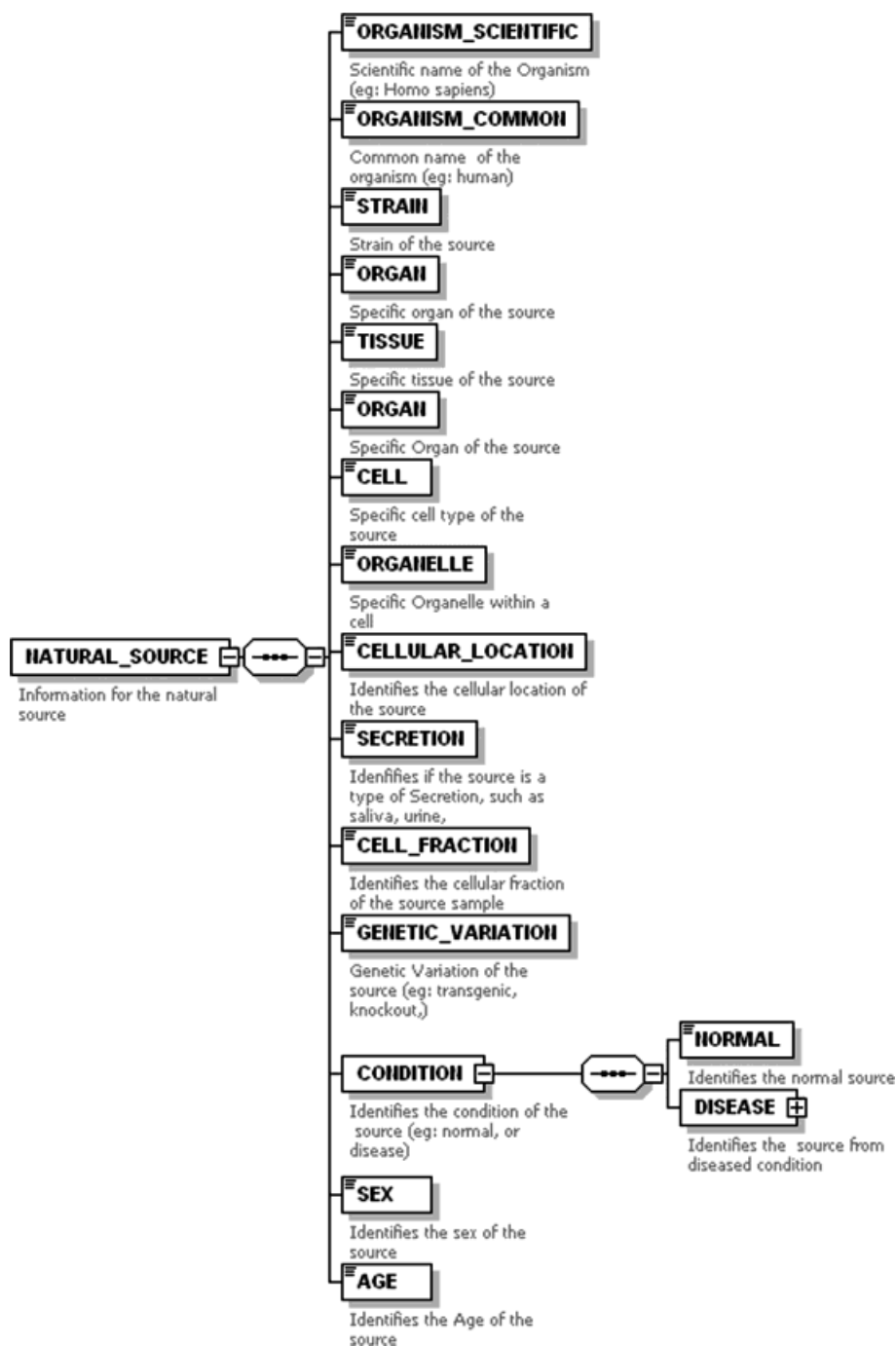
Figure 2. Schematic diagram for the natural source.

sensitive information. Two newer XML schema languages, the W3C's XML schema definition language [17] and ISO's RELAX NG [18], address these shortcomings of DTDs. Therefore, we plan to specify the TWODML schema using one or both of these languages. Our application will also include one or more extensible style sheet language (XSL) style sheets. These style sheets will enable electrophoresis XML data to be transformed into other useful formats such as HTML (hyper text markup language) and SVG.

2.4 Data repository or electrophoresis data warehouse

Many specialized database systems exist throughout the world that focus on 2-D electrophoresis data, with SWISS-2DPAGE being one of the major sources [19]. However, these run in isolation and rarely cooperate with each other, and there is not a reliable way to allow comparisons to be made between data sets. This raises the
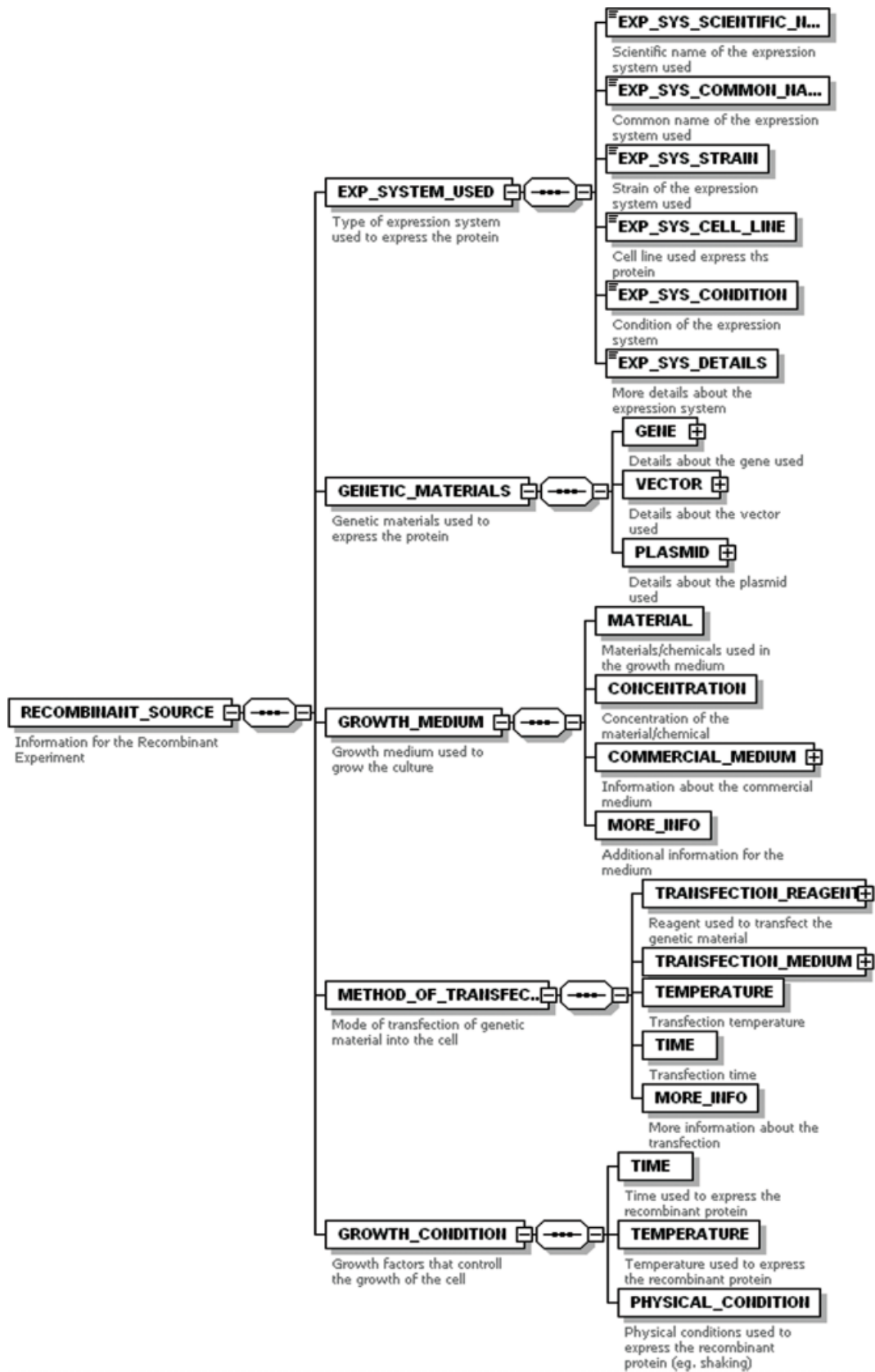
Figure 3. Schematic diagram for the recombinant source.

Figure 4. A part of recombinant source XML model data.

problem of data exchange. Providing interoperability among these databases is important to the successful operation of an organization. Improvements in productivity will be gained if the systems can be integrated – that is, made to cooperate with each other – to support global applications accessing multiple databases. Hence, a common repository such as the Electrophoresis Data

Repository (Fig. 5) proposed here for electrophoresis data would be ideal. This process model, described in the following sections, uses TWODML as a format for query results. Thus, information from the repository can be shared with any other system supporting TWODML. Also, XML tools may be used to make the query results human-viewable.
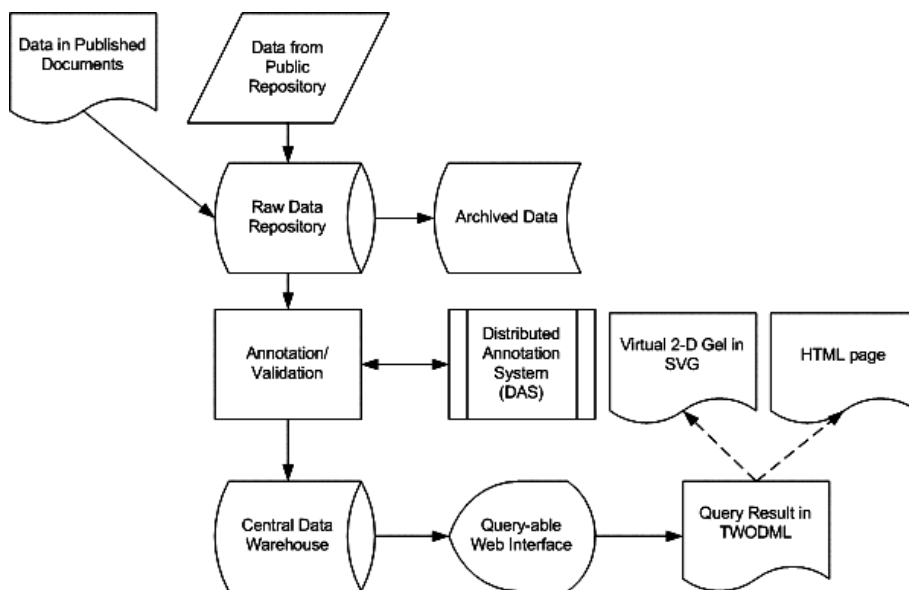
Figure 5. Process model for 2-D data warehouse.

2.4.1 Data source

One of the main objectives of the Electrophoresis Data Repository is to provide the community with detailed information about a given protein of interest, including qualitative and quantitative properties. This requires more comprehensive data for each of the data items and data groups. One way to achieve this is to encourage the electrophoresis community to deposit their experimental data, so this will facilitate the entry of electrophoresis data and associated information through a web-based common repository. It is clear that having to supply such detail for every single parameter in an experiment data set can be a highly onerous task for data submitters. If a public repository is to encourage submis-sion of experimental data, the designers of such databases must strive to reduce the amount of manual labor required of submitters. Software tools accompanying electronic repositories must provide the equivalent of GeneBank's SEQUIN (http://www.ncbi.nlm.nih.gov/Sequib), BankIt (http://www.ncbi.nlm.nih.gov/BankIt), or Gene Expression Omnibus (GEO: http://www.ncbi.nlm. nih.gov/geo/). Nonetheless, if an electrophoresis data repository has sufficient data in enough categories then it will be possible to query the required data. Also, a parallel effort has to be established to gather electrophoresis data from the published articles and integrate them into the repository. This knowledge component of a resource is usually held in scientific natural language as text. Extracting 2-D experimental information could also be done using information retrieval (IR). IR is the field of computer science that deals with the processing of documents containing free text, so that they can be rapidly retrieved based on keywords specified in a user's query [20]. Data derived form the public repository as well as form the mined data can be stored in a raw data repository. Raw data should be archived in a standard format, so as to ensure the data integrity, originality, and traceability.

2.4.2 Data annotation/validation

Electrophoresis data are complex and heterogeneous, that should be annotated with the community's understanding about the data. Each data point and its value needs to be examined or validated for its correctness and completeness. The annotation of data elements also requires that all of the related data records within a file are consistent and properly integrated across the group of files. The first phase will be largely automated; annotation resources that will be routinely consulted to provide a complete range of updated, annotated information. Continuous annotation of electrophoresis data will improve the interoperability of cross-platform data sets. Since we are adopting TWODML, the annotation of electrophoresis

data is also extendable to Distributed Annotation System (DAS). DAS is a client-server system in which a single client integrates information from multiple servers to retrieve and integrate dispersed proteomics data [21]. Since DAS adopts the integration through XML, well-annotated electrophoresis data can be integrated from various specialized resources through TWODML.

2.4.3 Data management and archiving

The volume of electrophoresis data is growing at a nearly exponential rate and this poses a problem in terms of data management, scalability, and performance. Building of the database structure is the first step towards the structured recording of electrophoresis data in a relational database. This consists of precisely defined data fields and precisely defined relationships between them represented by links between the tables. Electrophoresis data are complex to model and there are many different types of data presenting numerous relationships. Data models are the logical structures used to represent a collection of entries and their underlying one-to-one, one-to-many, and many-to-many relationships. The main motivation for creating electrophoresis data models is usually to be able to implement them within database management systems, usually as a relational database management system (*e.g.*, ORACLE, SQL Server, SYBSASE, MySQL, *etc.*). This electrophoresis data model corresponds to a way of organizing the pertinent values obtained on measurement of an electrophoresis experiment. This database is being modeled to handle the heterogeneous data from various external data sources. New types of proteomics data emerge regularly and this raises the need for updating the whole data semantics, and integrating the sources of information that were formally independent. Data analysis generates new data that also have to be modeled and integrated.

2.5 Query-able Web interface

A customizable query interface should permit complex queries that include the most commonly required data for electrophoresis. The individual or grouped data should be able to be downloaded in TWODML format.

2.5.1 Data distribution

Data are accessed intensively and exchanged very often by users on the Internet. Users can gather their data in TWODML format through a web interface that can be queried. This is useful for data exchange between other systems compatible with TWODML. On the other hand, users might also want to view the data. In this case, the

data must be presented in a user-friendly format. An obvious choice is HTML since it is supported by all Internet browsers. Another, perhaps more compelling format useful for human viewing is SVG – a standard XML format for describing 2-D graphics. Unlike vanilla HTML, SVG drawings are dynamic, interactive, and may even be animated. HTML and/or SVG can be obtained from TWODML query results using XSLT.

2.5.2 A proposed system

Standards are adopted by: (i) single vendor controls a large enough portion of the market to make his product the market standard; (ii) community agrees on an available standard specification; (iii) group of volunteers representing interested parties work in an open process to create a standard; (iv) government agency such as the National Institute of Standards and Technology (NIST) coordinates the creation of consensus standards. The NIST has an established record for the development and certification of validated and standardized databases. NIST is currently developing a program to meet some of the needs of the proteomics communities. As an exploratory model system for 2-D PAGE data standards, we experimented with the 2-D PAGE data of mitochondrial proteins [22]. Mitochondrion is an ideal target for global 2-D PAGE because of its manageable level of complexity.

2.5.3 Theoretical model

As an initial attempt at standardization, implementation of a 2-D PAGE reference based upon the theoretical calculation of isoelectric point (pI) and molecular weight for each mitochondrial protein (both mitochondrial and nuclear encoded) sequences has been carried out. A customizable interface was developed to permit complex queries that include the name of the protein, tissue, mitochondrial compartment, chromosome number where relevant, molecular weight range, pI range, and keywords (Fig. 6). The query results, along with the protein name, pI and molecular weight are presented as an intermediate selection, with the protein name linked to the theoretical virtual 2-D PAGE

of that protein. The virtual 2-D PAGE shows the query protein's mobility on a virtual 2-D gel, based upon the pI and molecular weight as calculated from the protein sequence (Fig. 7). Each 2-D spot's protein name is linked to detailed information about the protein (Fig. 8). The information presented on the detail page includes: SWISS-PROT information, description, cellular location, key words, tissue, cellular function, similarity with other proteins, gene name, synonyms, chromosomal location, and protein sequence information such as the amino acid length, theoretical pI, and theoretical molecular weight. External database links are also presented on this page. The protein sequence can be highlighted using a mouse-over option that provides annotation such as the mitochondrial localization signal, variant information, and other protein sequence details. The protein sequence of interest can also be used to search the journal references, the SWISS-PROT site for related proteins, or against the Protein Data Bank sequence for the related 3-D structures.

2.5.4 Experimental model

We plan to extend the 2-D PAGE option of our theoretical model into an experimental model. In order to be able to search a 2-D PAGE database with selectable query options for an individual protein, it is necessary to have comprehensive data for each of the individual categories. Currently, we are experimenting to mine the 2-D PAGE data from the published literature and incorporate them into our model data warehouse. The issue of effectively describing the data elements for spot intensity will be critical and should thus be carefully considered. The spot intensity, direct image file data, should be considered separately from protein concentration reporting, derived data from the image intensity. Physical calibration standards will probably be required to be captured in order to permit one to associate protein concentration with image intensity. The issue of recording spot intensity data has been worked out in other disciplines, such as the much more complicated X-ray crystallography case, where image plates or detectors capture X-ray diffraction reflections. Applications of these or similar algorithms would

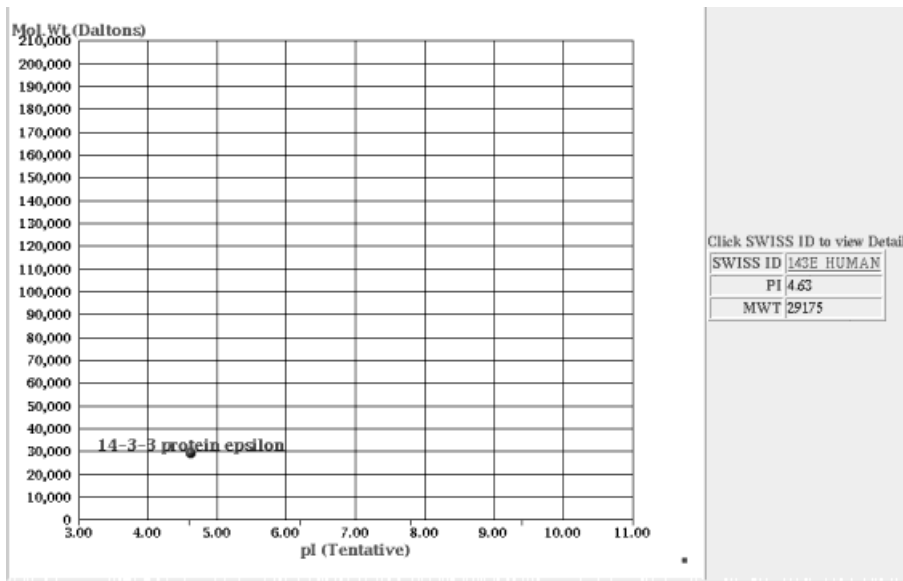**Figure 6.** Query-able web interface for mitochondrial 2-D (theoretical) data.

Figure 7. Virtual 2-D gel image for 14-3-3 protein.

Figure 8. Example of the detailed information page from a search for protein 14-3-3.

seem prudent. We are also in the process of exploring physical 2-D PAGE standards and associated algorithms to dynamically detect and measure 2-D spots from diverse 2-D gels. Our mitochondrial 2-D PAGE information is available from the NIST mitochondrial protein web site at: http://bioinfo.nist.gov:8080/examples/servlets/index.html.

A community effort towards a similar goal for 2-D PAGE meta data is urgently needed. Individual research consortia, such as PSI, HUPO, or the Electrophoresis Society are in a good position to explore informatics approaches that will work within a consortium, but the time is ripe for the formation of *ad hoc* groups at various proteomics meetings, and it is hoped that this article acts as the catalyst for such collaboration. *Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?"* – T. S. Elliott

The contents of this article do not necessarily reflect the views or policies of NIST. Any mention of commercial products within this article is for information only and does not imply recommendation or endorsement by NIST. The World Wide Web pages are provided as a public service by NIST. With the exception of material marked as copyright, information presented on these pages is considered public information and may be distributed or copied. Use of appropriate byline/photo/image credits is requested.

3 References

[1] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al.*, *Science* 2001, *291*, 1304–1351.

[2] Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gestetland, R., Walters, L., *Science* 1998, *282*, 682–689.

[3] Collins, F. S., McKusick, V. A., *J. Am. Med. Assoc.* 2001, *285*, 540–544.

[4] Molloy, M. P., *Nat. Biotechnol.* 2003, *21*, 597.

[5] Frank, R., Hargreaves, R., *Nat. Rev. Drug Discov.* 2003, *2*, 566–580.

[6] Petricoin, E. F., Zoon, K. C., Kohn, E. C., Barrett, J. C., Liotta, L. A., *Nat. Rev. Drug Discov.* 2002, *1*, 683–695.

[7] Renfrey, S., Featherstone, J., *Nat. Rev. Drug Discov.* 2002, *1*, 175–176.

[8] Verrills, N. M., Kavallaris, M., *Curr. Opin. Mol. Ther.* 2003, *5*, 258–265.

[9] Srinivas, P. R., Verma, M., Zhao, Y., Srivastava, S., *Clin. Chem.* 2002, *48*, 1160–1169.

[10] Goldstein, D., *Manag. Care Interface* 2003, *16*, 68–69.

[11] Stupka, E., *Curr. Opin. Mol. Ther.* 2002, *4*, 265–274.

[12] Campagne, F., *Nature* 2002, *418*, 125.

[13] Jones, J., Preston, H., *Top. Health Inf. Manage.* 2000, *21*, 45–54.

[14] Achard, F., Vaysseix, G., Barillot, E., *Bioinformatics* 2001, *17*, 115–125.

[15] Mackenzie, D., *Science* 1998, *280*, 1840–1841.

[16] World Wide Web Consortium, Extensible Markup Language (XML) 1.0 (second edition), W3C recommendation, 6 October 2000, http://www.w3.org/TR/REC-xml

[17] World Wide Web Consortium, XML Schema Part 1: Structures, W3C recommendation, 2 May 2001, http://www.w3.org/TR/xmlschema-1/

[18] Organization for the Advancement of Structured Information Systems (OASIS), RELAX NG Specification, committee specification, 3 December 2001, http://relaxng.org/spec.html

[19] Hoogland, C., Sanchez, J. C., Tonella, L., Binz, P. A., Bairoch, A., Hochstrasser, D. F., Appel, R. D., *Nucleic Acids Res.* 2000, *28*, 286–288.

[20] Nadkarni, P. M., *Pharmacogenomics J.* 2002, *2*, 96–102.

[21] Stein, L. D., *Nat. Rev. Genet.* 2003, *4*, 337–345.

[22] Verma, M., Kagan, J., Sidransky, D., Srivastava, S., *Nature Rev. Cancer* 2003, *3*, 789–795.