LBL-29385
UC-408

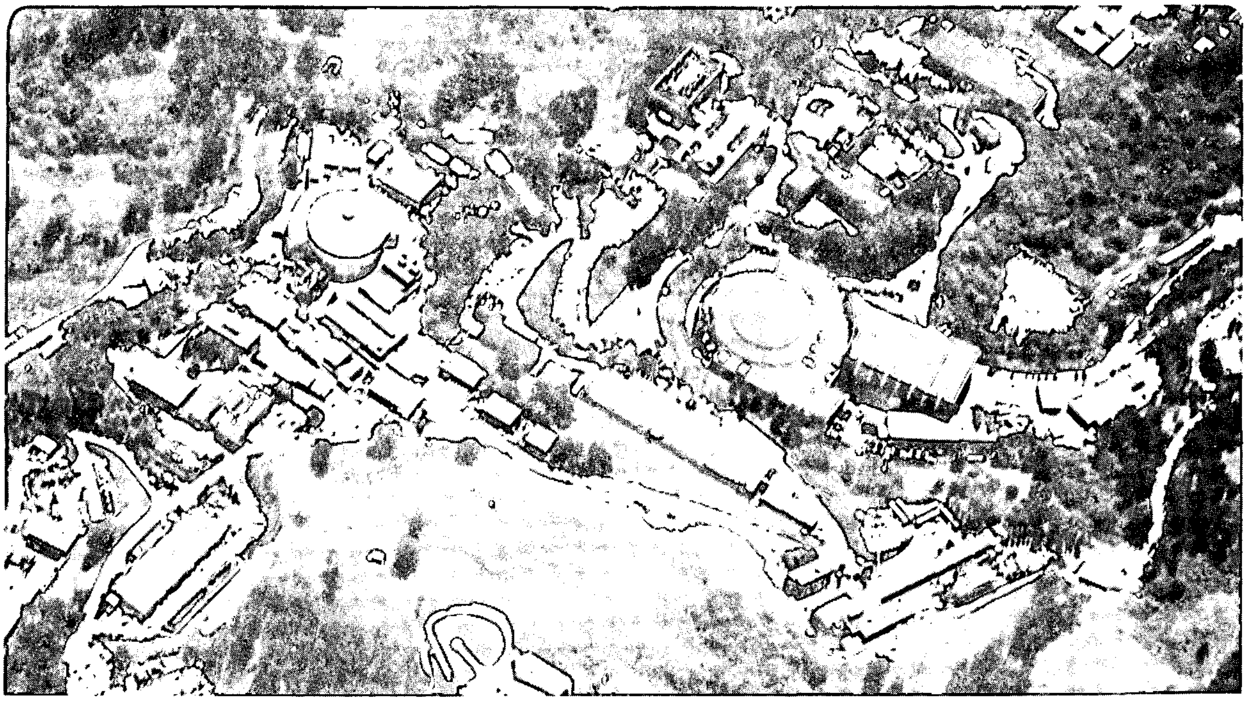
Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

Information and Computing Sciences Division

Data Interchange Standards for Biotechnology: Issues and Alternatives

J.L. McCarthy

June 1990



Prepared for the U.S. Department of Energy under Contract Number DE-AC03-76SF00098.

DISCLAIMER

LBL-29385

Data Interchange Standards for Biotechnology: Issues and Alternatives

John L. McCarthy

Computing Science Research & Development
Information & Computing Sciences Division
Lawrence Berkeley Laboratory
1 Cyclotron Road
Berkeley, California 94720

June 1990

This work was supported by the National Center for Biotechnical Information, National Library of Medicine, National Institutes of Health, and the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

Data Interchange Standards for Biotechnology: Issues and Alternatives

John L. McCarthy
Computer Science Research and Development Department
Information and Computing Sciences Division
Lawrence Berkeley Laboratory
Bldg 50B-room 3238
Berkeley, CA 94720

JLMcCarthy@LBL.GOV (arpanet)
JLMcCarthy@LBL (bitnet)
415/486-5307

Prepared for

National Center for Biotechnical Information

National Library of Medicine

June, 1990

This work was supported by the National Center for Biotechnical Information, National Library of Medicine, National Institutes of Health, and the U.S. Department of Energy under Contract Number DE-AC03-76SF00098.

DISCLAIMER: This document was prepared as an account of work sponsored by the United States Government. Neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial products, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California and shall not be used for advertising or product endorsement purposes.

APPENDIXES

1. PURPOSE OF THIS REPORT

As several workshops and reports have noted, the current lack of standards for biotechnical data (including syntax, semantics, and nomenclature) is a substantial barrier to scientific progress. In order to facilitate data exchange and interoperability of data analysis programs, the biotechnical community needs to adopt some basic guidelines about the form and content of its data and data transactions.

In the biotechnical sciences, as in other fields, output from one database or computer program may become input for another. Data interchange standards are thus desirable because they facilitate the reciprocal complementarity of input and output, databases and programs. They encourage modularity and interoperability of software tools for data manipulation and analysis. In doing so, they can reduce the over-all costs of software development while increasing scientific productivity.

This report outlines a framework for discussion of what aspects of biotechnical information might be good candidates for guidelines or standards, what existing data exchange standards might be appropriate building blocks upon which to build, and what procedural mechanisms might be appropriate for adoption of such guidelines or standards. It builds on experience from other scientific communities which have already benefitted from development of discipline-specific data exchange standards.

The report is organized as follows: Section 2 introduces and distinguishes different facets of data exchange that ought to be considered for guidelines or standards. Section 3 briefly outlines some major types of biotechnical information that already are being shared or need to be shared. Section 4 proposes various criteria for deciding what approaches to biotechnical data interchange standards might be most appropriate. Section 5 describes major standards organizations that currently formulate data exchange standards, and section 6 outlines pros and cons for several existing standard syntax choices which the biotechnical community might look to as a starting point. Section 7 concludes the report with some brief recommendations about what next steps ought to be taken, as a stimulus for further discussion. Four appendices provide (A) example templates for summarizing standard information about attributes of biotechnical entities; (B) a comparative example of ASN.1 versus TSDN syntax; (C) a selected bibliography of pertinent written materials; and (D) a list of organizations and individuals who might play a role in this effort.

2. DIFFERENT ASPECTS OF DATA EXCHANGE STANDARDS

Interchange standards can pertain to different aspects of data that are in fact independent (orthogonal) dimensions. It may be easy to standardize some aspects but not others. Blurring of the distinctions can lead to unnecessary confusion and disputes. The purpose of this introductory section is to distinguish different aspects of data that may be standardized.

An Analogy and an Example

Data exchange is a special instance of human communication in general, and written language in particular. For written language, just as for data, we can distinguish between semantics, syntax, nomenclature, and formatting issues.

If two or more people wish to communicate, they must first agree on a common language. There are substantial benefits to using a widespread language which is already familiar to many people and which has many associated tools (e.g., dictionaries, thesauri, etc.), versus making up a new language or using a relatively obscure one. There are also arguments in favor of using a alphabetic language, rather than a pictographic one, such as ease of representation and interpretation in electronic form.

Within a given language, one needs to observe certain syntactical rules (e.g., sentences with subjects and predicates, placement of nouns, verbs, adjectives, adverbs, and so on). In addition to syntax, our written communications also must conform to certain semantic constraints -- what types of information go with what other types of information. For example, the statement "the dog flies green" is syntactically correct, but semantically meaningless. Nomenclature issues have to do with what we call certain types of things -- e.g., whether we use a standard set of names for countries. Finally, written communications may be encoded and formatted in particular ways -- e.g., a particular type font within lines of a book with one inch margins. Note that we may have different standards or guidelines that may apply to each of these different levels independently.

Moving from analogy to a concrete data example, let us take a simple instance from the Human Gene Mapping Library MAP database. The following part of one record from that database pertains to a single locus, the gene SOD1.

```
SYMBOL = SOD1;
UNIQUE.ID = LM0147;
LOCUS.TYPE = G;
OLD.SYMBOL = IPO-A, SODS, SOD-A;
XREF.GENBANK = J02947, K00065, M13267, X01780, X01781, X01782, X01783;
XREF.LIT = H4850, H5957, H7465, H7104;
MAP.LOC.STRUCT;
 MAP.LOC.TEXT = 21q22.1;
 REGION.TOP = 20.99000;
```

```
  REGION.BOT = 20.98510;
 DATE.ADDED = 07/17/89;
  MARKER.NAME = superoxide dismutase 1, soluble;
  E.C.NUMBER = 1.15.1.1;
  MIM.NUMBER = 14745;
```

This example illustrates a number of different aspects of data interchange standardization: specification of entities and attributes; syntax structure and symbols; forms of names and data values; controlled vocabularies; and data transactions. Let us examine each of these in turn.

Specification of Data Objects

One of the first questions an interchange standard must address is what type of data objects are to be interchanged. Any given interchange may contain data for one or more instances of one or more attributes pertaining to one or more types of biotechnical entities. In the above example, we have a collection of attributes (SYMBOL, UNIQUE.ID, LOCUS.TYPE, etc.) which pertain to a single instance of a particular type of biotechnical entity (a chromosomal locus). As this example also illustrates, several other standardization issues relate closely to what types objects are interchanged.

Attributes and Their Components

What attributes (fields) should comprise information for each type of entity? Should particular attributes be mandatory for all entities (e.g., name and unique identifier) or for certain types of entities? What aspects of attributes themselves ought to be standardized (e.g., names, synonyms, measurement units or domains of data values, etc.)? Appendix A contains an example template for summarizing standard information about attributes.

Data Structure and Representation

Should exchange data objects be restricted to simple, single-valued attributes (as in relational tables) or can they include complex (and possibly nested) data structures? Can a single standard exchange file or transmission contain information for different types of entities? Some have suggested that the SQL (Structured Query Language) and relational tables are sufficient, while others argue that we need to be able to represent complex objects in a direct and understandable manner (such as in the simple HGML example above).

The HGML example illustrates how certain fields (attributes) may contain something other than a single string or numeric value (as would be the case in a simple relational table). Note also how OLD.SYMBOL, XREF.GENBANK, and XREF.LIT all contain multiple values (which are separated by commas). Furthermore, the MAP.LOC.STRUCT attribute is a "repeating group" of other attributes (MAP.LOC.TEXT, REGION.TOP, and REGION.BOT), which itself may occur one or more times. Whether and how to permit representation of such nested data structures is a key aspect of data interchange standardization.

Syntax Structure and Symbols

The generic type of syntax used in the above example is called "tag-value" because each data value is "tagged" by a preceding identifier for each attribute (e.g., SYMBOL and LOCUS.TYPE). An alternative syntax might drop the tags and specify sequentially ocurring fields (separated by some symbol, such as ";" or "@") in an external "codebook." Use of internally imbedded tags is most appropriate for data that may be irregular in occurrence, length, or sequential order. Fixed field or fixed sequence formats are most appropriate for data that maintains much the same form from instance to instance over time.

Syntax symbols are used to separate substructures from one another -- e.g., tags from values (here by the equal sign "="), and tag-value pairs from one another (here by semi-colons ";"). Such symbols can be changed (such as "=" to ":" and ";" to "@") independent of syntax structure.

Forms of Names and Data Values

Data interchange standards often impose constraints on the form of tags or data values, such as the following:

- fixed length or variable length up to a specified maximum
- prohibition of imbedded blanks, especially for tags
- restriction to a limited character set, such as alphanumeric ASCII

For example, HGML tags must begin with a letter or $, do not distinguish upper and lower case, must be fewer than 17 characters and cannot contain special characters other than period, hyphen, $, or underscore. Constraints for HGML data values vary from field to field. Text fields can contain up to 32,000 characters; others can contain only numbers or a controlled vocabulary.

Controlled Vocabularies

Data exchange standards that employ tags may require that such tags belong to a controlled vocabulary, perhaps with provision for use of aliases or synonyms as part of the controlled domain of tags. Standards may also impose controlled vocabulary constraints on values of certain attributes, particularly those that are coded or which are "foreign keys" that link one type of information to another (e.g., the code "G" for LOCUS.TYPE or GenBank sequence numbers for XREF.GENBANK in our HGML example).

Data Transactions

Beyond simple data transmission, an increasing number of data exchange standards have begun to specify certain types of transactions, such as debiting a specified account or changing a specified attribute value for a particular record (e.g., "change SYMBOL = SOD-A to SYMBOL = SOD1 in MAP record LM0147"). The banking and purchasing communities have invested substantially in development of data transaction standards such as X12 (Electronic Data Interchange or EDI) [X12, 1988].

3. MAJOR TYPES OF BIOTECHNICAL INFORMATION

What types of biotechnical information might be covered by data exchange standards or guidelines? If we simply consider different types of information available in public databases, the spectrum of entities is broad, and the associated sets of attributes are extensive. The list below outlines some of the major types of biotechnical entities that ought to be considered in any standardization effort.

Sequences. Nucleic acid sequences for DNA and RNA, amino acid sequences for proteins and polypeptides, and associated annotations such as "feature tables" that relate functional components to specific locations within a given sequence.

Maps. Cytogenetic maps, genetic linkage maps, radiation breakage linkage maps, physical maps, and other maps which picture different types of sites along individual chromosomes.

Genes and Other Loci. Data about specific locations on a chromosome, including functions and features.

Reagents. Information pertaining to restriction enzymes, probes, and other reagents

Strains. Information about specific strains of cells, tissue, plants, animals, etc.

Allelles and Polymorphisms. Genotypic and phenotypic information about variants of a particular gene or locus.

Physical Structure. Atomic, crystallographic, and other structural aspects of proteins, DNA, and other macro-molecules

Bibliographic Information. Citations, with or without abstracts or full text of articles

Experimental Data. Input parameters, instrument readings, and analyzed results

Images. Digitized images from auto-radiograms, confocal microscopy, etc.

This brief, incomplete list suggests that any standardization efforts will need to be quite general, flexible, and extensible (i.e., easily extended to accommodate new types of information). Rather than go into any of these specific types of information in more detail, let us consider what general requirements a data exchange standards effort needs to address for all of them.

4. PRIMARY REQUIREMENTS AND CONSIDERATIONS

Regardless of the specific types of information, there are several general types of requirements that need to be weighed in considering data exchange standards. For any particular discipline or type of data, the relative importance of these considerations will vary. Indeed, some considerations may conflict, such as human-readability versus transmission efficiency, and we must choose which is more important. At this point, the major considerations for biotechnical data interchange appear to be as follows:

Ease of Use

The more transparent a data interchange format is, the more likely its success. End users should not have to be aware of it, except perhaps to specify input or output as BIF (Biotechnical Interchange Format), much as they might specify the RTF text interchange format to convert from one text processing program to another (e.g., WORD to WORDPERFECT).

Ease of Implementation

An interchange format also needs to be as simple as possible for application program and database developers to implement. Transparent access for end users in turn requires that those who provide databases and analysis software implement facilities to read and write data in the interchange format. They will not do so unless the interchange format is broadly accepted and relatively simple to implement. One thing that can make implementation simpler is availability of subroutine libraries to translate from the interchange format to specific programming language or database data structures and vice-versa.

Semantic Completeness

Many of the types of biotechnical information outlined above have a rich semantic content which needs to be fully captured by any data interchange format. This includes representation of relationships, classes, inheritance, aggregation and generalization hierarchies.

Object Representation

The most straight-forward, concise, and easy-to-understand form of representation currently available for the rich semantics of much biotechnical data are object-oriented data structures, rather than sets of relational tables. Objects may have attributes which have single values (as in the relational data model) or multiple values. Moreover, attributes may consist of multiply-occurring sets of other attributes (repeating groups), with nesting and recursion.

Component Independence

As in databases, distinct concepts should be cleanly separated rather than bundled together. A data exchange format should encourage users to decompose data into discrete, standardized semantic components. For example, measurement units and values should appear in separate components rather than a single string (e.g., "value=20; units=cm" rather than "20cm") in order to reduce ambiguity and facilitate computer processing. Data that has been broken down into highly differentiated components can always be recombined. It is much more difficult, and sometimes impossible, to automatically extract finer detail from larger undifferentiated components.

Self-describing Syntax

In applications where the form and content of data being exchanged is well-understood and unlikely to change very much over time, description of the data can be separate from the data itself. Most early interchange formats began with (and many still use) fixed fields to represent different types of information. Both sender and receiver must use the same "codebook" that describes the location and possible contents of each field.

In fields which are rapidly changing, such as biotechnology, it is highly desirable to have a more flexible, "self-describing" syntax, in which variable length fields are identified by mnemonic tags or labels that specify their contents explicitly. This minimizes the need for "out of band" communications, such as separate codebooks, and thus reduces opportunities for error and misunderstanding.

Extensibility

Another aspect of rapidly changing fields such as biotechnology, is that data exchange standards must be extensible. That is, it must be easy to add new types of entities, attributes, permissible values, and so-on, with minimal disruption of prior versions of the standards.

Human Readability

Although data may be encoded in various ways for efficiency of transmission and storage, it is desirable that human beings be able to read and edit some version of the data interchange format in terms of standard ASCII characters and text editing tools. This will not only facilitate debugging of software that uses the interchange format, but also to permit people to verify the contents at different stages of exchange (e.g., from database to analysis program, between analysis modules, in temporary work files, etc.)

Use of Existing Standards

Use of existing data exchange standards can simplify the standards development process because existing standards usually specify certain structural and syntactic constraints, and they may have associated tools to facilitate getting data into and out of a standard exchange format. The question is whether the constraints imposed by existing standards are appropriate for biotechnical data (the next section addresses this issue).

Interchange standards have been developed for a wide variety of different data. Some are specific to certain disciplines (e.g., astronomy) while others are intended as generic "building blocks" and have been used for diverse types of data. The biotechnical community can choose whether to begin with one of the generic interchange standards and build on it, or to invent something "from scratch," using insights from previous efforts in other fields.

5. DATA EXCHANGE STANDARDS ORGANIZATIONS

This section outlines major organizations that are concerned with data interchange standards, from international organizations whose scope is very broad, to national and discipline-specific groups.

The **International Standards Organization** (ISO) is headquartered in Geneva, Switzerland. Its work is carried out by different standards committees (SC), which are in turn broken up into Working Groups (WG). Members of ISO Standards Committees and Working Groups represent constituent national standards organizations, rather than individuals or corporations. There are a number of ongoing ISO data interchange standards efforts in different SC's and WG's.

Another international group whose mission specifically concerns scientific data sharing and standards is **CODATA** (Committee on Data for Science and Technology), which was created by the International Council of Scientific Unions (ICSU) in 1966. In addition to holding biannual meetings, CODATA's Task Groups have developed data exchange standards in a number of different disciplinary areas, including biotechnology. In 1987, the CODATA Task Group on Coordination of Protein Sequence Data Banks published "A standardized format for sequence data exchange" [George, Mewes, and Kihara, 1987]. In 1988, CODATA established a new Commission on the Terminology and Nomenclature of Biology, which is currently attempting to coordinate conventions and standards developed by individual biologically oriented unions and affiliates of ICSU -- to promote standard terminology where it exists and to identify areas in need of further standardization.

Many United States members of ISO are representatives from committees of the **American National Standards Institute** (ANSI). Like ISO, ANSI has hundreds of different committees and working groups organized in a hierarchical structure. X3 is the parent committee on Information Processing Systems. X3T2 is the committee responsible for generic data interchange standards, such as ASN.1. In the U.S., EDI and Product Definition Exchange Standard efforts are under the auspices of ANSI Committee X12 (Electronic Business Data Interchange). Library, information retrieval and publishing standards fall under the National Information Standards Organization (NISO), also known as ANSI Committee Z39. Membership on ANSI committees can be either individual or corporate (though it is usually the latter). Members pay annual dues and contribute efforts on a volunteer basis.

The **American Society for Testing and Materials** (ASTM) is another standards organization that has recently been extending its work into data interchange standards as well as the more traditional areas that its name suggests. The most relevant ASTM committees are E-31 (Computerized Systems), and its subcommittee E-31.12 (Medical Informatics). The latter has concerned itself primarily with patient care records, but its scope also includes "biomedical research." In November, 1989, ASTM began a new development activity on standard interchange formats for computerized chemical data [ASTM, 1990].

The National Institute for Science and Technology (NIST, formerly the National Bureau of Standards) plays an active role on many ANSI and ASTM committees, as well as their international counterparts. NIST often takes the lead role in setting Federal Information Processing Standards (FIPS). Since federal procurements must conform to FIPS standards, that gives corporations a strong economic incentive to participate in standards efforts.

ANSI X3T2 Data Interchange Committee and ISO SC22/WG11. Generic data interchange standards efforts reside primarily in one US national committee (ANSI X3T2), and its International Standards Organization working group counterpart, ISO SC22/WG11. Three of the leading candidate generic interchange standards discussed below (ASN.1, DDF, and TSDN) are under active review by X3T2 at this time. For a current membership list of individuals and organizations on X3T2, see [X3T2].

Other Interested Organizations. In addition to the traditional standards-setting organizations outlined above, there are several other groups that have expressed a strong interest in helping to create standards for biotechnical information. These include the National Center for Biotechnical Information (NCBI) at the National Library of Medicine (NLM), the Genome Projects Joint Informatics Task Force (Department of Energy/National Institutes of Health), and the recently formed international Human Genome Organization (HUGO).

6. DATA INTERCHANGE SYNTAX CHOICES

This section describes several specific data exchange syntax standards that might be candidates for biotechnical data. For each candidate, brief subsections describe its structure and syntax, current uses by other organizations, strengths, and limitations.

Data Descriptive File (ISO 8211)

Data Descriptive File (DDF) format originated from a U.S. Department of Energy Working Group during the 1970's. It was one of the first media and machine independent formats for interchanging information between computing systems. It became an ANSI standard early in the 1980's and an ISO TC97/SC15 international standard (ISO 8211) in 1984.

Structure and Syntax. The DDF format specifies file and data record descriptions via a modified tag-value format, in which tags are physically separated from the values to which they pertain. It was inspired, in part, by the MARC record format for bibliographic records developed by the library community. The following description is taken from [Gallagher, 1984]:

> ".... Record components may be elementary data elements,vectors, arrays, or hierarchies. The elementary elements may be character strings, bit strings, or various numeric forms.

> The DDF consists of a Data Descriptive Record (DDR), which describes the characteristics of each data field, followed by a sequence of Data Records (DR), which contain the actual data occurrences. The DDR and DR records have the same structure, consisting of "leader", "directory", and"data" portions. The "leader" is a sequence of 24 characters that gives the total record length in characters, codes for the level and type of the record, and information for reading the "directory" portion. The "directory" establishes integer "tags" that correspond to fields in the"data" portion of the record and gives starting positions and lengths for all such fields. For an interchange file consisting solely of fixed-length records containing only fixed-length data fields in which the DR's have identical leader and directory values, the leader and directory of the first DR apply to all subsequent DR's. The leader and directory of the subsequent DR's may be omitted.
>
> The "data descriptive area" of the DDR contains a "data descriptive field" for each of the "user data fields". Each data descriptive field associates a data name or a reserved word with each tag. The "data fields" of the "user data area" of the DR contain the user information to be interchanged. Each data field is an instance of the user data structure and data types defined by the DDR data descriptive field with the corresponding field tag. Data names in the DDR correspond to data values in the DR if and only if they have identical tags.
>
> The standard provides for three implementation levels from which users may choose depending on the complexity of their data structures. Level 1 supports multiple fields containing simple, unstructured character strings. Level 2 supports level 1 and also processes multiple fields containing structured user data comprising a variety of data types. The third level supports level 2 and hierarchical data structures."

Uses. DDF has been considered for use by some applications, but it is not currently being used for any to the knowledge of this author. X3T2 has a current work item to review and perhaps revise the DDF standard.

Strengths. Since ISO 8211 associates tags with character strings, it is somewhat more human-readable than ASN.1, though separation of those character strings from the data values makes human readability difficult at best.

Limitations. DDF files contain several layers of indirection, and hence can be rather difficult to understand. There are also some limitations (such as length of tags and depth of data structures) that could be detrimental. Since ISO 8211 has not been widely used, there are few tools, documentation, and people knowledgeable about its use.

CCIT X.409 and ASN.1 (ISO 8224, ISO 8225)

The most widely used generic data exchange standard in the last few years has been ASN.1 (Abstract Syntax Notation 1), an ANSI and ISO Standard that grew out of (or alongside of) the CCITT X.409 communications standard, and as a part of the more general Open Systems Interface (OSI) Basic Reference Model (ISO 8072).

As [Gallagher, 1984] notes,

> "The X.409 Presentation Syntax is more general than the DDF interchange format in that it is a "language" for defining general data structures rather than just a syntax for specifying columns and row occurrences of an underlying tabular structure. In addition to specifying definitional syntax, X.409 specifies a binary "encoding" for each defined data structure. Using this approach, information can be exchanged in two parts - the first part a character string that defines a specific data structure, and the second part a string of octets that is an encoding of a value of the defined data structure. X.409 was adopted by CCITT as a formal international Recommendation (i.e. Standard) in October 1984.
>
> The ISO ASN presentation syntax (ISO 8824) and basic encoding rules (ISO 8825) were originally intended to be identical to X.409, except that they are separated into two parts. A separate specification of language syntax separate from the basic encoding rules allows other alternative encoding rules to be defined over the same definitional syntax, e.g. a non-binary encoding....
>
> There have been some minor differences between X.409 and ASN -- in that ASN allows alternative syntax in several situations...."

Structure and Syntax. As Gallagher points out, the basic philosophy behind both X.409 and ASN is that there exist a number of elementary data types from which all other data structures can be defined. "Built-in" elementary data types include Boolean, Integer, Bit String, Octet String, and Null Value. Built-in "constructor" data types for defining more complex data structures include Sequence and Set. Data structures definable via X.409 or ASN syntax are those that can be constructed from these elementary data types using nested applications of constructor types such as sequences and sets. If desired, intermediate data structures may be tagged with user specified names, but such substantive meaning of tags must be communicated separately -- either in a common "codebook" or via a separate transmission (see example below).

> "Interchange of any data structure defined by X.409 [or ASN.1] syntax is accomplished using (Type/Length/Value) triples. The Type is a bit sequence that identifies a data type previously defined by the X.409 syntax, the Length is an integer that declares the length in bytes of a data occurrence, and the Value is the actual encoding of the data occurrence. Encoding rules for both elementary and constructor data types specify how each type is represented as a string of 8-bit octets. Any data structure definable via the X.409 syntax can then be encoded as a nested hierarchy of (type/length/value) triples, always with a linear representation as a string of 8-bit octets. It follows that any definable data structure can be exchanged on any medium capable of transporting these strings of octets."

Uses. ASN.1 was originally developed for specification of communications protocols, and that is still its field of greatest application, but it has subsequently been used by a much broader range of applications, including the ANSI X3H4 Standard for Information Resource Dictionary Systems.

Strengths. ASN.1 provides a simple, but rich set of basic constructs that are relatively easy to understand and capable of representing as much complexity as may be needed. It is a national and international standard that has been used in a growing number of different applications. With ASN.1, the syntax is relatively independent of encoding rules (i.e., the syntax can be represented by an ASCII "print form" as well as a binary encoding). There are tools (e.g., encoders and decoders), documentation (including several introductory tutorials), and people who understand how it works. In particular, NIST's OSI Toolkit reads and writes the ASN.1 ASCII print format [which is not yet part of the standard] as well as the ISO 8825 binary encoding [U.S. Dept of Commerce, 1989].

Limitations. The current ASN.1 standard does not specifically include a character string (ASCII) representation form of tags and values, although such a form is used in examples and may be considered for inclusion as part of the standard in the future. Although the standard separates the abstract syntax from the encoding, the syntax currently assumes that tags are encoded, rather than possibly remaining as human-readable character strings. Since the tags are encoded, senders and recipients have to agree on predefined meanings of all tags and data structures, or else to define them in a separate transmission.

Transfer Syntax Description Notation (NASA)

TSDN is a more recently developed generic data interchange standard. It has not yet been adopted as either an ANSI or ISO standard, but is still undergoing review by those bodies as well as final revisions by its main sponsors, NASA and the European Space Agency. Fred Billingsley, TSDN's primary author at the Jet Propulsion Laboratory in Pasadena, expects it will go out for initial public review from X3T2 late in 1990.

TSDN was developed to describe not only data for exchange, but also the vast archives of existing space science data (much of it from telemetry) which reside in a bewildering variety of arcane formats on tape and other media.

Structure and Syntax. Like ASN.1, TSDN is a language rather than a format *per se*, but it supports a richer variety of capabilities. According to a recent paper by Billingsley,

"The basic approach can be simply stated: transmit whatever is required, in whatever form desired, but describe the transmission in ASCII characters according to recognized description syntax - TRUTH IN LABELING.

.... It permits a sender to describe the transferred information and to send this description separately or as an integral part of the transfer file. It permits the description of both character and bit field information in fixed- (without delimiters) or variable-width (delimited) fields or subfields. It further permits the identification of fields and subfields by arbitrarily long names and labels which serve to give meaning to the data. In addition, it provides for the definition and labeling of complex structures and commutated (coded) data."

Uses. TSDN is still under development, so it has only been used with test data sets thus far. TSDN documentation includes a number of extensive examples of how it can be used to represent various types of application data structures -- some of them quite complex.

Strengths. Unlike ASN.1, TSDN descriptions themselves are part of the proposed standard. The proposed standard provides that such descriptions can be sent with the data or in a separate transmission. Prototype tools are under development to provide basic services, such as retrieval of a named data element from a particular file in a user-specified format. TSDN has been developed under the auspices of the international Consultative Committee on Space Data Sciences (CCSDS), with funding from NASA Code EC. It is under consideration by ANSI committee X3T2, the same committee responsible for DDF and ASN.1. Since it is still not cast in concrete, there are still opportunities to improve TSDN before it becomes a national or international standard (perhaps in 1991).

Limitations. TSDN is not yet a fully developed national or international standard. It does not yet have a body of associated tools and experts, let alone a track record of successful use. The TSDN syntax itself seems a bit more clumsy, baroque, and non-intuitive than ASN.1. It has a somewhat more rigid and old-fashioned syntax of its own (with parentheses delimited parameter lists that must be entered in a specified order), but that may be because it is trying to do more, and because it is designed for efficient processing more than formal human-readable description.

Related Conventions and Standards

Other possibilities have been suggested to facilitate exchange of biotechnical data. We will only discuss them briefly here, but some may warrant more detailed consideration.

In 1984, a Format Subcommittee of the new CODATA Task Group on the Coordination of Protein Sequence Data Banks adopted a set of general recommendations which developed into a detailed formal standard two years later [George, Mewes, and Kihara, 1987]. This format was designed to facilitate conversion to and from existing sequence formats as well as computer processing. It is based on a general, context-independent free format so that new types of information can be added without interference to existing software. Its components are similar to earlier GenBank and Protein Information Resource (PIR) formats. Since it is defined in a formal Backus-Naur Form (BNF) specification, it could also be framed in terms of a general purpose interchange definition language such as ASN.1 or TSDN.

The current GenBank Transaction Protocol goes beyond the earlier GenBank and CODATA sequence formats in that it attempts to address transactions as well as individual data entities (see the last paragraph in section 2 above). It is already being used to exchange sequence information, and it could be adapted to cover other types of data. This approach has been criticized, however, on the grounds that the protocol uses a "homegrown" language to specify entities, attributes, and other aspects of the data, as well as transactions. "For us in the biological databases to design a language seems at best unnecessary, and at worst we are likely to make a bad job of it.... It seems to us that the language and the software to process it are daunting enough to prevent it ever catching on as a standard."[Cameron, 1989].

An alternative proposal is to base everything on relational data structures and the standard Structured Query Language, SQL [Cameron, 1989]. The problem with that approach, as noted above, is that relational representation of biotechnical information is far from intuitive or straightforward. Anyone exchanging data would have to agree on a common relational schema, or at least understand how source and target schemas differ for any given interchange. Furthermore, there is as yet no standard for transmission of relational tables and their descriptions, although ANSI X3H4 has proposed an interchange format for information resource dictionary systems based on ASN.1 that can include relational tables and entity-relationship information.

Another approach that has been suggested is to use an object-oriented language such as IDL (Interface Description Language) that already has tools for translating to and from standard programming language (e.g., C) data structures [Pecherer, 1989]. While such an approach might be expedient in the short term, it probably would not be sufficiently general for the longer term -- especially if based on a particular proprietary language.

Finally, it may be worth noting that an increasing number of organizations, particularly in the publications industry, have been adopting the MARC record format for data interchange. Originally developed by the Library of Congress for bibliographic records, MARC is a "tag-value" format similar to DDF (which it helped inspire) that includes complex objects, repeating values, and so on.

Why not wait for emergence of a *de facto* standard from commercial vendors? A number of de facto data interchange standards (DIF for spreadsheet data, dBase for microcomputer database files, etc.) have become commonly accepted because of the popularity of particular commercial products. In the case of the biotechnical community, however, it seems unlikely that one organization is likely to dominate the diverse market sufficiently to put a whole set of *de facto* standards in place -- except perhaps for special niches such as the GenBank transaction format.

7. WHERE SHOULD WE GO FROM HERE?

Although the choice of a particular approach to developing biotechnical data exchange standards is not obvious, it does seem clear that the biotechnical community would be wise to choose an existing standard language (such as ASN.1 or TSDN) within which to develop definitions of specific entities, attributes, and so on. TSDN looks like a good choice in the long run, but it is not yet a standard and does not yet have a body of associated tools. ASN.1 can be used immediately to develop such definitions (as Jim Ostell has demonstrated). Although the ASN.1 standard does not yet include ASCII representation and transmission of the definitions themselves, NIST's OSI Toolkit does read and write the ASN.1 ASCII "print format."

Since ASN.1 appears to be readily translatable into TSDN, and since it is more human-readable, the community could begin by using ASN.1 to build consensus about definition of objects and attributes of interest -- putting aside questions of specific syntax for the time being. It also could begin to build systematic lists of attributes such as begun in Appendix A. At the same time, biologists could begin to participate on X3T2 in order to track and influence the further development of both ASN.1 and TSDN.

What does seem clear is the need for serious standards development efforts -- i.e., national and international committees under existing standards organizations. One problem here is choice of appropriate umbrella national and international standards organization(s). ANSI, ASTM, ISO, CODATA, or a professional society such as ACS are all potential candidates. The DOE/NIH Joint Informatics Task Force, HUGO, and NCBI have all expressed a strong interest in standards development activities. Although the conventional mechanisms for standards development can seem painfully slow, past experience suggests that they may be the best we can hope for. Others who have tried to short-circuit the process -- such as the computer aided software engineering (CASE) community -- have often taken as much if not more time and becoming part of the conventional standards process in the end anyway.

A. EXAMPLE FORMATS FOR INFORMATION ABOUT ATTRIBUTES

This appendix suggests a template for systematic cataloguing of standard summary information about attributes of biotechnical entities. This type of information can be very helpful in deciding which attributes ought to be mandatory or optional for all types of entities or specific types of entities. It also helps focus attention on what aspects of each attribute ought to be subject to standardization, such as names and aliases, measurement units, value domains (i.e., numeric ranges or controlled vocabularies), and so on.

This template is not intended to be comprehensive. Its purpose is to serve as a starting point and strawman for discussion of what a full template and list of attributes ought to include. Each attribute needs to be systematically described in terms of its name, information content, permissible values, and so on. Based on experience from other fields, the following types of information are desirable for minimal characterization of individual attributes:

Name

Description

Occurrence (Single or Multiple)

Data Type (e.g., Real, Integer, Date, Category set, String)

Units or name of restricted vocabulary category set

Requirement (Mandatory/Recommended/Optional)

Constraints (e.g., range for numeric information)

Standard (name of relevant standard or organization)

Example(s)

Comments

It is easier to read and evaluate a list of attributes if such information is displayed in a systematic way. Two alternative display formats are illustrated below; the first is a summary table, while the second is a "tag:value" format.

Attr. Name/Description	Occ	Type	Units	Req.	Constrs/Cmnts
organism	1	C	MeSH	M	
scientific name of organism					
common name	M	S	text	O	
strain	1	C	*	M	*set by organism committee
cell line	1	C	ATCC		

Name:	organism
Description:	Scientific name of organism
Occurrence:	M
Data type:	string
Standard:	taxonomic nomenclature group
Requirement:	required
Example:	Sus scrofa
Comments:	

Name:	common name
Description:	
Occurrence:	M
Data type:	string
Standard:	Nomenclature committee of the organism
Requirement:	optional
Example:	pig, swine

Name:	strain name
Description:	Strain of the above organism, when applicable.
Occurrence:	1
Data type:	string
Standard:	Nomenclature committee of the organism
Requirement:	required
Comments:	In humans, it may be useful to add an attribute for race

Name:	cell line
Description:	Name of cell line from which the element was derived.
Occurrence:	1
Data type:	string
Standard:	ATCC or IMR nomenclature conventions
Requirement:	required when applicable

B. A COMPARISON OF ASN.1 AND TSDN

The following comparison of ASN.1 and TSDN representation is based on a personnel record example that both ASN.1 and TSDN developers have used, for the following reasons:

(1) it does not require specialized knowledge of a subject area, which could be a problem for different biotech subspecialties as well as people who are experts in computer data and interchange formats but not biology.

(2) it does not distract biotech specialists with questions about the substantive content -- as opposed to the exchange format *per se*

(3) it has become something of a standard example

The logical structure of the data example, as given in Annex E of the ASN.1 documentation (ISO 8824) is as follows:

Name:	John P Smith
Title:	Director
Employee Number:	51
Date of Hire:	17 September 1971
Name of Spouse:	Mary T Smith
Number of Children:	2
Child Information	
Name:	Ralph T Smith
Date of Birth:	11 November 1957
Child Information	
Name:	Susan B Jones
Date of Birth:	17 July 1959

The formal ASN.1 description, as shown in Annex E, is as follows:

```
Personnel Record ::= [APPLICATION 0]  IMPLICIT SET
  {                    Name              ,
    title              [0] VisibleString ,
    number             EmployeeNumber    ,
    dateOfHire         [1] Date          ,
    nameOfSpouse       [2] Name          ,
    children           [3] IMPLICIT
                       SEQUENCE OF
                       ChildInformation
                       DEFAULT { }       }
    ChildInformation ::= SET
      {                         Name              ,
        dateOfBirth             [0] Date}

    Name ::= [APPLICATION 1]  IMPLICIT SEQUENCE
      {givenName            VisibleString,
        initial             VisibleString,
        familyName          VisibleString}

    EmployeeNumber ::= [APPLICATION 2] IMPLICIT INTEGER

    Date ::= [APPLICATION 3] IMPLICIT VisibleString –YYYYMMDD
```

The physical instantiation of John Smith's personnel record is formally described in the ASN.1 8824 Annex E in the following "print format:"

```
  {                   {givenName "John", initial "P",familyName "Smith"} ,
    title             "Director"                                          ,
    number            51                                                  ,
    dateOfHire        "19710917"                                          ,
    nameOfSpouse      {givenName "Mary", initial "T", familyName "Smith"},
    children
    {{{givenName "Ralph", initial "T", familyName "Smith"},
      dateOfBirth "19571111"                          },
    {{givenName "Susan", initial "B", familyName "Jones"},
       dateOfBirth "19590717"   } }    }
```

The following TSDN description is taken from Billingsley's paper:
"We will visualize a physical data record in which the data fields are separated by commas and terminated with an exclamation point, thus:

John,P,Smith,Director,51,19710917,Mary,T,Smith,2,Ralph,T,Smith,19571111,Susan,B,Jones,19590717!

We use the capitalization forms of 8824 (TSDN is case-insensitive):

```
MCAI = TSDN,1989-11-01 ;             /* Identifies TSDN date of issue */
UID = ASN.1 Personnel Record Example ;  /* User-defined identification   */
+++                          /* Completes the General Section */
TYPE = givenName,(A,)        ;
TYPE = initial,(A,)         ;  /* Allows multiple initials */
TYPE = familyName,(A,)       ;
TYPE = title,(A,)           ;
TYPE = number,(I,)          ;  /* Allows multi-digit employee number */
TYPE = year,I4,("1900".."2001") ;  /* Note constraint on numeric ranges */
TYPE = month,I2,("01".."12") ;
TYPE = day,I2,("01".."31") ;
TYPE = numberOfChildren,(I,) ;
OBJECT = PersonnelFile,SET,+(PersonnelRecord) ;
OBJECT = PersonnelRecord,SEQUENCE,(Name,title,number,dateOfHire,
  nameOfSpouse,numberOfChildren,numberOfChildren(ChildInformation)) ;
OBJECT = Name,SEQUENCE,(givenName,initial,familyName) ;
OBJECT = nameOfSpouse,SEQUENCE,Name;
OBJECT = ChildInformation,SEQUENCE,(Name,dateOfBirth) ;
OBJECT = dateOfHire,SEQUENCE,date
OBJECT = dateOfBirth,SEQUENCE,date
OBJECT = date,SEQUENCE,(year,month,day) ;
FILEID =PersonnelFile,NONE,<fileName>; /*Makes file available to system*/
***            /* Completes the Detail Section... optional data records follow */
John,P,Smith,Director,51,19710917,Mary,T,Smith,2,Ralph,T,Smith,19571111,
Susan,B,Jones,19590717!
@@@                          /* Completes the TSDN Module    */
```

NOTE - statements may be made to visually resemble the ASN.1 construction by lining up the components:

```
    OBJECT = PersonnelRecord,SEQUENCE,
        (
            Name,
            title,
            number,
            dateOfHire,
            nameOfSpouse,
            numberOfChildren,
             numberOfChildren(ChildInformation)
        );
    OBJECT = Name,SEQUENCE,
        (
            givenName,
            initial,
            familyName
        );
    OBJECT = nameOfSpouse,SEQUENCE,
        (
            givenName,
            initial,
            familyName
        );
    OBJECT = ChildInformation,SEQUENCE,
        (
            Name,
            dateOfBirth
        );
```

*[***Does TSDN require specification of number of ChildInformation?]*

C REFERENCES AND SELECTED BIBLIOGRAPHY

Overviews, Tutorials, and Tools

Chappell, D. (1986) "A Tutorial on Abstract Syntax Notation One (ASN.I)" Omnicom Information Service Report 25, Omnicom, Inc., Vienna, Virginia, December 1986.

Zajaczkowski, J.A. (1987) "An Introduction to the CCITT/ISO standard on transfer syntax and notation" *British Telecom Technology Journal* (October 1987)

Billingsley, F. and J. Johnson (1990) "Language Requirements for Understanding Retrieved Data" *IEEE Symposium on Mass Storage Systems* [forthcoming]

Billingsley, F. C., J. Johnson, E. Greenberg and M. MacMedan. (1989) "Facilitating Information Transfer in the Eos Era." *IEEE Transactions on Geoscience and Remote Sensing* 27(2): 117-124.

Gallagher, L. (1984) "Data Interchange Forms" National Bureau of Standards.

U.S. Department of Commerce, National Institute for Science and Technology (1989) "OSIkit Tools from NIST," June 1989.

W. Stallings (1987) *Handbook of Computer-Communications Standards — The Open Systems Inter-connection (OSI) Model and OSI-Related Standards.* Volume 1, Macmillan, New York, 1987

Carl F. Cargill (1989), *Information Technology Standardization: Theory, Process and Organization* (Digital Press, 1989)

Biotechnical Data Exchange

Cameron, G. (1989) "The GenBank Transaction Protocol -- The EMBL View"

Cinkowsky, M. (1989) "The GenBank Transaction Protocol and the Proposed NCBI Data Exchange Format"

GenBank. (1989) "Introduction to the transaction protocol used to update the GenBank database"

George, D. G., Mewes, H.W., and Kihara, H. (1987) "A standardized format for sequence data exchange," Protein Sequences and Analysis 1 (1987), 27-39.

Ostell, J. (1989). ASN.1 Definitions of Sequence Data.

Pecherer, R. (1989). "IDL - Interface Description Language: Notes and Comments." :

Generic Standards Documents

NOTE: ISO documents can be obtained from OMNICOM. 501 Church St., N.E., Suite 304, Vienna, VA 22180

CCITT (1988) Specification of Abstract Syntax Notation One (ASN.l). CCITT Recommendation X.208, 1988.

CCITT (1988) Specification of Basic Encoding Rulcs for ASN.l. CCITT Recommendation ,Y.209, 1988.

Consultative Committee for Space Data Systems. (1989). Transfer Syntax Description Notation.

Consultative Committee for Space Data Systems. (1989). Transfer Syntax Description Notation Fundamental Support Level.

Consultative Committee for Space Data Systems. (1989). Transfer Syntax Description Notation Supporting Document.

ISO 8824:1987(E), "Information processing systems - Open Systerns Interconnection- Specification of Abstract Syntax Notation One (ASN.l)", International Organization for Standardization, Switzerland, 11/l5/87

ISO 8825: 1987(E), "Information processing systems - Open Systems Interconnection- Specification of Basic Encoding Rules for Abstract Syntax Notation One (ASN.l)", International Organization for Standardization, Switzerland, 11/15/87

ISO 8823:1988(E), "Information processing systems - Open Systems Interconnection- Connection oriented presentation protocol specification", International Organization for Standardization, Switzerland, 8/15/88

ISO/IEC 8824/DAD 1, "Information processing systems - Open Systems Interconnection - Specification of Abstract Syntax Notation One (ASN.l)", ADDENDUM 1: ASN.l Extensions, International Organization for Standardization, Switzerland, 06/09/88

ISO/IEC 8825/DAD 1, "Information processing systems - Open Systems Interconnection - Specification of Basic Encoding Rules for Abstract Syntax Notation One (ASN.l)", ADDENDUM 1: ASN.l Extensions, International Organization for Standardization, Switzerland 06/09/88

ISO 8211, "Data Descriptive File for Information Interchange", International Organization for Standardization, Switzerland, 9/84

Discipline-Specific Standards

American Society for Testing and Materials. (1990). to Attendees at the Planning Meeting for a Proposed New ASTM Voluntary Standards Development Activity on Standard Interchange Formats for Compuerized Chemical Information and Other Interested Individuals.

ANSI Acredited Standards Committee X12 (1988) "An Introduction to Electronic Data Interchange"

McDonald, R. S. and P. A. Wilks. (1988). "JCAMP-DX: A Standard Form for Exchange of Infrared Spectra in Computer Readable Form." Applied Spectroscopy. 42(1):

Wilhoit, R. C. and A. Maczynski. (1987). COdataSTAndardThermodynamics: Rules for Preparing a COSTAT Message for Transmitting Thermodynamic Data.

Related Standards Documents

ANSI X3T2 - Data Interchange, Draft Minutes of Eleventh Meeting: 24-27 January, 1989, San Diego, CA

DAF: Working Document on ASN-l, Version 2, October 1988, Red Bank, NJ

X3T2, Membership List, January, 1990

ISO. Information Processing - Open System Interconnection - Basic Reference Model. ISO International Standard 8072, 1983.

D. ORGANIZATIONS AND PEOPLE

There are a number of organizations and people who may have an interest in participating in data interchange standards and development efforts. A partial, initial list appears below, grouped roughly by type of organization.

NLM/NCBI: Ostell, Benson, Lipman, Peter Karp, others???

Other NIH: Guyer (Genome Center); Micah Krichevsky(Microbial Strains)

Other Agencies: John Wooley, Bob Robbins(NSF); Diane Hinton(HHMF); Donna Maglott, Lois Blaine(ATCC); Jeff Schmaltz(DOE),

Other biotechnical computing people: Temple Smith, Eric Lander, Tom Marr, Jamie Carbonell, Maynard Olson, Will Gillete, Nate Goodman, Gio Wiederhold; Glen Evans (Salk), George Bell, Tim Hunkapillar (Cal Tech), John Deveroux, George Church, David Adler (NeXT Genome Machine), Mark Tuttle (NLM Metathesaurus), Gene Myers, Bruce Schatz (AZ)

Other Computer Science Data Interchange people: Patrick Powell(MN)

DOE Labs: Rob Pecherer, Deborah Nelson (LANL), Ross Overbeek(ANL), Dave Thurman(PNL), Rowland Johnson, Elbert Branscomb (LLNL)

Major biotechnical databases: Mike Cinkowsky (LANL/GenBank), Al Hillyard (Mouse/Bar Harbor), Graham Cameron (EMBL), Sarai? (DDBJ); Peter Pearson, Dick Lucier (Hopkins/Genome Database), Ken Kidd, Mike Mador or Mark Cavanaugh (Yale/HGML); Benjamin or Vicki Nichols (CAS), David George (PIR), Tom Koetzle (Brookhaven/Protein Structure), Steve Bryant (ICRF), Akira Tsugita (Japan Protein Databank)

Overseas collaborators: Chris Rawlings, Martin Bishop (UK), Sarai (Japan), Akira Tsugita (Science University of Tokyo)

Other Genomes: David Mount, Stan Letovsky, Mary Berlin (E.Coli); Bob Mortimer, David Bottstein (Yeast); Sydney Brenner (C.Elegans)

Standards Groups: Fred Billingsley (X3T2/TSDN), Mark Hamilton, Richard Foote (X3T2/ASN.1), Al Brooks (X3T2/ISO8211), Dana Marks (X3H4?IBM Repository), Alan Goldfine (X3H4/NIST/FIPS), Paul Peters (NISO, ALA, X3) [NY Public Library 212 930-0720]

Other Interchange Experts: Nick Roussopoulos (Hopkins), Dana Marks (IBM), Harvard Holmes (LBL)

Commercial Vendors: Bob Gross (TextCo), David Beech (Oracle); ____?(Intelligenetics), Frances Lewitter (BBN),

[this list needs more analysis & tools people -- e.g., IBI, Meyers]