Lawrence Berkeley National Laboratory
Recent Work

Title
Computers and Biotechnical Terminology: Accessing and Integrating Data from Diverse Sources

Permalink
https://escholarship.org/uc/item/1ss0b1cx

Author
McCarthy, J.L.

Publication Date
1990-03-01

eScholarship.org Powered by the California Digital Library
University of California

UC 605

LBL-28752

Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

Information and Computing Sciences Division

Presented at the Second Pacific Chemical Society Conference, Honolulu, Hawaii, December 17-22, 1989, and to be published in the Nomenclature of Biotechnology Proceedings

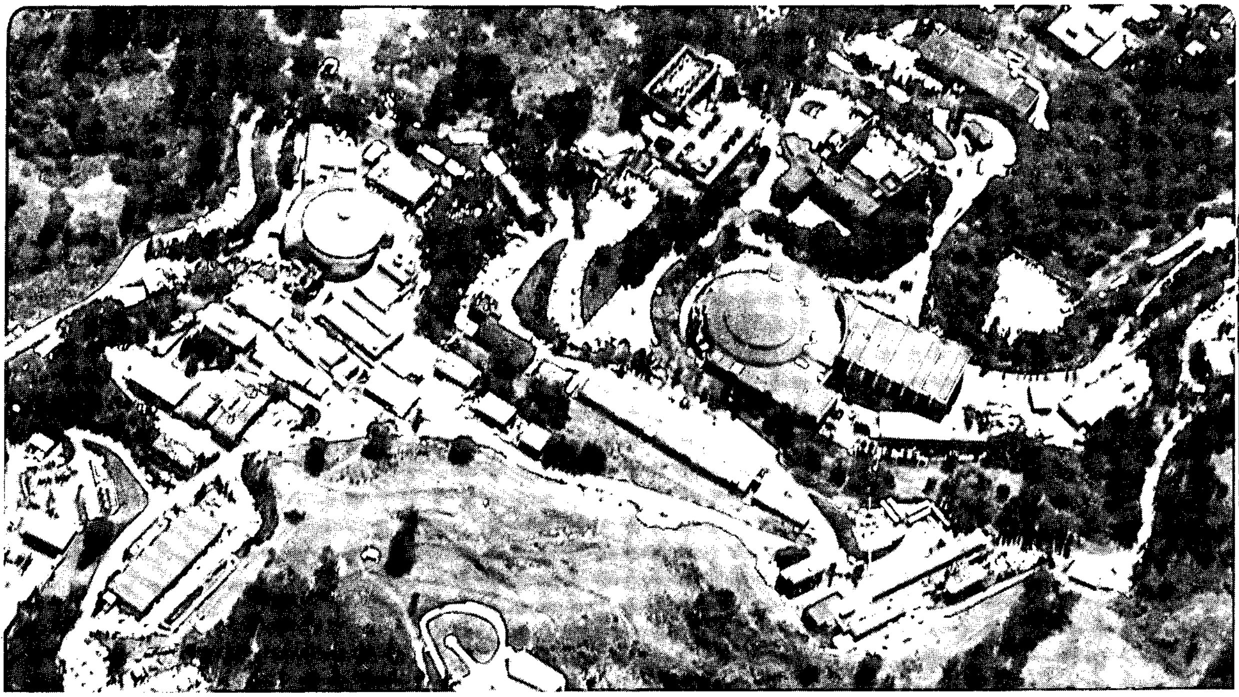
Computers and Biotechnical Terminology: Accessing and Integrating Data from Diverse Sources

J.L. McCarthy

March 1990

For Reference

Not to be taken from this room



LBL-28752
Copy 1
Bldg. 50 Library.

Prepared for 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. While this document is believed to contain correct information, 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 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 product, 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.

LBL-28752

COMPUTERS AND BIOTECHNICAL TERMINOLOGY: ACCESSING AND INTEGRATING DATA FROM DIVERSE SOURCES

John L. McCarthy

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

March 1990

In The Nomenclature of Biotechnology Proceedings, 2nd Pacific Chemical Society Conference, Hawaii, December 1989.

This work was supported by the Director, Office of Health and Environmental Research Program, and Human Genome Program, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

COMPUTERS AND BIOTECHNICAL TERMINOLOGY: ACCESSING AND INTEGRATING DATA FROM DIVERSE SOURCES

John L. McCarthy

Information and Computing Sciences Division
Lawrence Berkeley Laboratory
University of California
Berkeley, CA 94720 U.S.A.

Abstract: Biotechnical information on computers presents special terminology challenges for locating and integrating computer data from diverse sources and levels. Standard nomenclature problems such as synonyms, homographs, and name changes are compounded when they arise on computers for names of databases, entities, and variables, as well as data values from multiple independent sources.

At the same time, computers make possible the development of new tools to meet these challenges. The data thesaurus is one such tool that helps integrate different types of data. It provides a systematic framework within which both people and computer programs can reconcile terminology from diverse, autonomous databases. Data thesauri also can help guide evolution of international standards for nomenclature and classification.

INTRODUCTION

Nomenclature is a serious obstacle to accessing and integrating our growing wealth of biotechnical data. Computers, which have fueled the data explosion, have also brought a host of new terminology challenges. This paper examines how insights and tools from computer science can help us understand and address those problems.

In order to discuss the special terminology challenges of biotechnical data, we will need to use some other terms (some might say jargon) from database management and computer science. Exhibit 1 gives a small subset of data that will introduce these terms as well as illustrate different aspects of data representation on computers.

Entities, Attributes, Values, and Domains

An ***entity*** is some "thing" about which we have data. The data in exhibit 1 pertain to a polypeptide entity named "interferon eicosapeptide." ***Attributes*** are characteristics or descriptors that pertain to an entity. In a specific computer representation, they sometimes are called fields or columns. Exhibit 1 shows three attributes: Registry, Name, and Sequence. Instances of an attribute for an entity (e.g., 79113-16-9) are called ***values***. ***Domains*** are classes or sets of permissible values for one or more attributes (such as a particular numeric range, or a set of codes -- e.g., for amino acid residues).

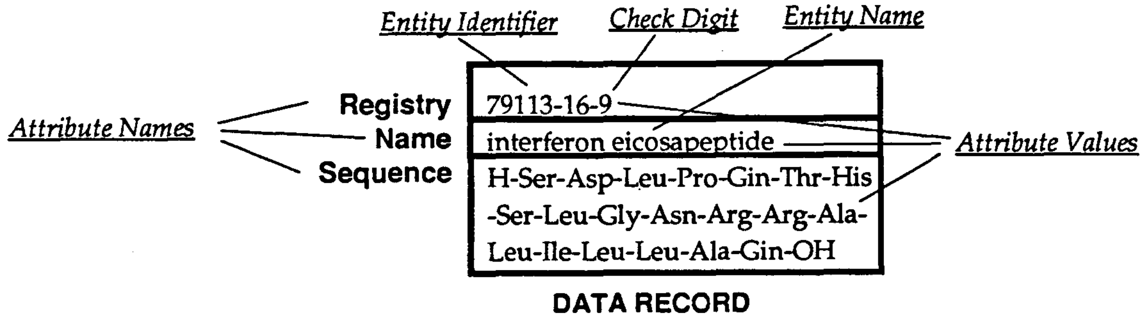


Exhibit 1: Computer Science Terms describe data components of a simple example

The terms ***type*** and ***instance*** refer respectively to general and particular aspects of other concepts -- such as entity. Thus we might say that exhibit 1 contains data for one specific *instance* of a "polypeptide" *type* of entity.

Identifiers, Names , and Data

An ***identifier*** is an attribute (sometimes called a "primary key") which unambiguously and unalterably identifies a particular instance of an entity -- such as the Chemical Abstracts Service Registration Number in exhibit 1 or a social security number for people. In order to guarantee that identifiers do not change, some experts have suggested that they have to be "meaningless" (essentially an arbitrary accession number) because any meaningful components of an identifier are subject to change [1]. To reduce transcription errors, identifiers also may contain a ***check digit*** (e.g., "9" in exhibit 1) [2; 3].

At the same time, people need ***names*** that are easy to remember -- attributes that may contain some meaningful content which can be used in place of identifiers. In the biosciences, things are frequently named to reflect biological function (such as uv123 for an ultraviolet radiation sensitive locus). Some entities may have more than one name (e.g., common name vs. formal name). Many papers in this symposium concern construction and assignment of names for specific types of entities such as proteins, carbohydrates, enzymes, interferons, monoclonal antibodies, genes, drugs, and so on.

Data is the generic term used to describe attribute values, such as the amino acid sequence in exhibit 1. Attribute values for a particular entity are frequently grouped together in physical or logical ***data record***. Values for each attribute may be constrained to a generic or special ***data type*** (e.g., integer, string, date, amino acid sequence) and to either single or multiple values (e.g., Registry vs. Sequence). In some cases (not shown in this example), attributes may themselves be complex ***data structures*** -- that is, constructs made up of other attributes. For example, an audit trail structure might consist of pairs of change dates and identifiers for persons who made each change.

Data and Metadata

Metadata is data about data, such as attribute names and types, that can be used to describe and control data values to which they pertain [4]. Metadata entities (e.g., "Attributes") in turn have their own meta-attributes, such as names, labels, synonyms, and so on. For example, the labels "Registry," "Name," and "Sequence" in exhibit 1 are metadata for attributes that describe a whole database of peptides.

DATA RETRIEVAL AND INTEGRATION REQUIREMENTS

Many people who deal with biotechnology need to *locate* relevant information and *integrate* data from different sources. These requirements have implications for the form and substance of both data and metadata. Experience from information retrieval, database management, and computer language design suggests a number of issues and guidelines.

Locating Relevant Information

There are two basic ways of using metadata to find data of interest. On the one hand, one may directly request information by specified attribute names and data values (e.g., "find Sequence values that match "ACTGGTAGCCTAAG" with fewer than 2 percent errors"). On the other hand, one may wish to browse the data or metadata at a particular level of detail (e.g., "show what attributes are available for human genes").

As the number of entities, attributes, or data values becomes large, they become difficult to locate or browse unless grouped into hierarchical classes [5]. As in a library subject catalog or thesaurus, any metadata item may participate in multiple classifications. For example, interferons can be classified as proteins (chemicals) as well as anti-infectives (drugs). The logical structure is not a simple hierarchy, but rather a directed acyclic graph, since a given node may have multiple higher level nodes.

Combining Data from Different Sources and Levels

Users want to not only browse and retrieve specified subsets of data, but also to combine and analyze data from different sources and levels. For example, one might want to identify known oncogenes on a particular human chromosome (from the Human Gene Mapping Library database), and then retrieve any available sequence information on such genes (from GenBank), along with information on related protein products (from the PIR database). Exhibit 2 (adapted from [6]) outlines current major levels of biotechnical information and databases at each of those levels.

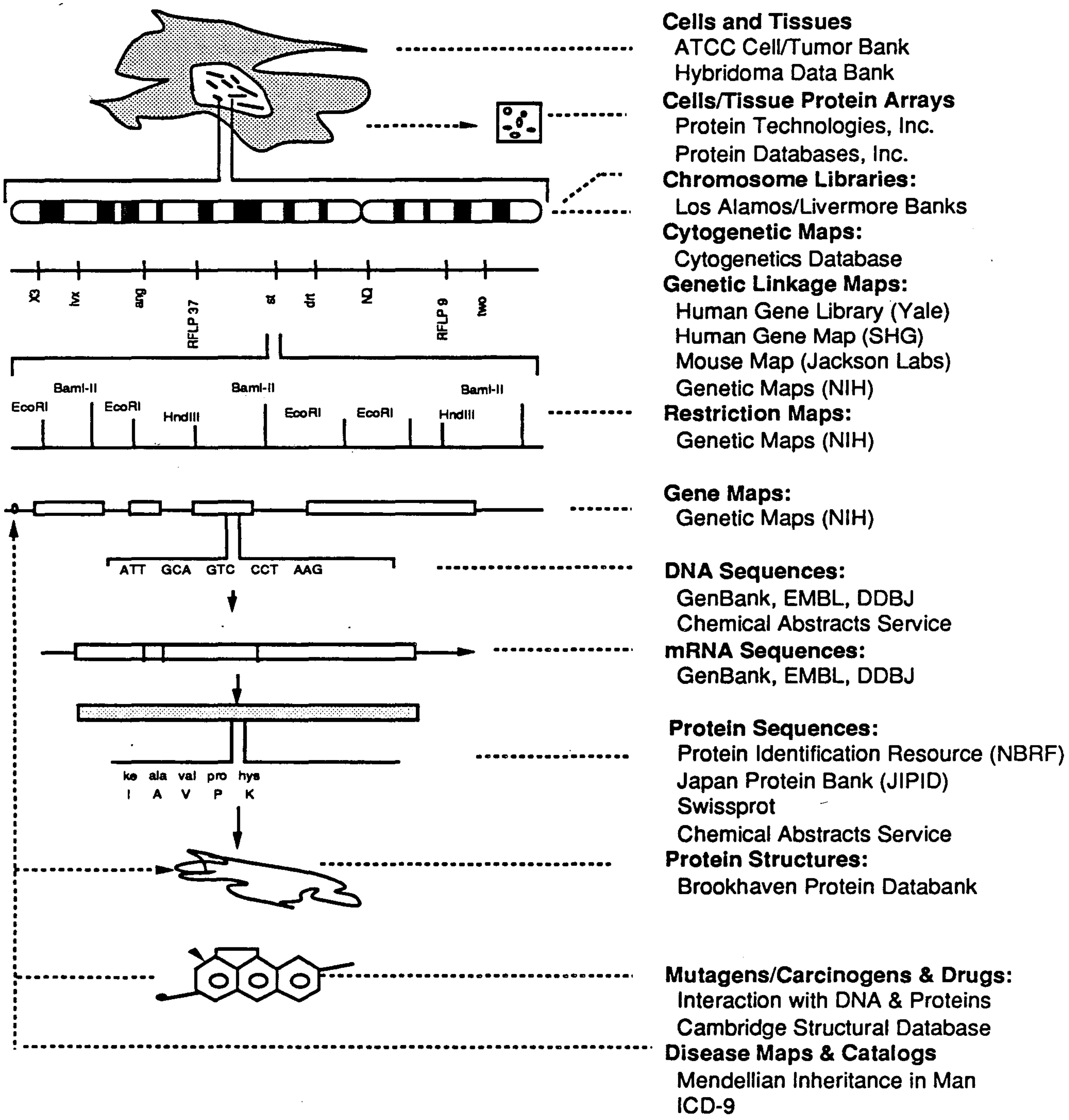


Exhibit 2: Different Types of Biological Data come from diverse databases
(adapted from "Biomedical Databases in a Universal Hierarchy of Nature" [6])

There are two basic ways to connect or link information between different databases -- direct and indirect. The most direct approach is to include identifiers from one database as attributes (foreign keys) in the other database. For example, HGML locus attributes include identifiers for literature citations (another HGML database), as well as GenBank sequences and OMIM (On-line Mendelian Inheritance in Man) entries. The indirect connection method is to match entities from different databases on the basis of attributes whose domains they share in common (e.g., gene names).

We can connect information from different databases indirectly if and only if (1) the entities in question share at least one attribute in common, and (2) the values for that common attribute can be resolved to the same domain. For each database, we must ascertain which attribute(s) may contain such information. For each such attribute, we need to know whether its domain is textual or numeric. Are values constrained to a controlled vocabulary or numeric range? What are the measurement units, if any? Finally, is the format of the attribute values fixed or"tagged" with their attribute name? Depending on the answers, it may or may not be possible to combine the information.

Consider, for example, information about physical position on a human chromosome, as represented in two major databases, GenBank and HGML, and summarized in exhibit 3.

	GenBank	HGML
Attribute Name	Location, Origin	Region
Attribute Values Domain	base pairs from Origin, relation to landmark	chromosome band
Measurement Units or Encoding	Kilobases	band number
Format/Syntax	embedded text	tag = value

Exhibit 3: Combining Data from Different Sources involves several comparability issues

Attribute Name. In GenBank, the position of an entity such as a gene is contained in a "feature table component" (attribute) whose name is Location, with measurements relative to the Origin (another attribute) of the sequence. In HGML, the comparable data element (attribute) is called Region, and its index is called Map.loc.

Value Domains. GenBank Locations are pairs of positive integers for "from" and "to/span"; in HGML Region values are cytogenetic band location codes (e.g., 21q22.3).

Measurement Type and Units. GenBank Locations are integer numbers of base pairs relative to the beginning of the sequence. The beginning position is called Origin, a text attribute with entries such as "7 bp upstream of SacI site." (Note that this "Origin" attribute is not the Origin of Replication Initiation, or ORI.) ISCN (International Standard Cytogenetic Nomenclature) band numbers in HGML are a decimal/ordinal code.

Format/Syntax. In GenBank, "from" and "to" components are separate items in a multiply ocurring data structure within the feature table. In HGML, ISCN code components (chromosome number, arm, major band, sub-band) are not separate.

Making and maintaining linkages between different databases can be done manually, on an *ad hoc* basis, or it can automated by computer software. In either case, linkage requires use of mutually agreed upon, controlled vocabularies (which can be augmented but not changed over time) for both attribute names and values, plus compatible measurement units and syntax. Terminology problems confound each aspect of data comparison.

TERMINOLOGY PROBLEMS OF COMPUTER DATA

Locating, comparing, and connecting data all face common terminology problems such as synonyms, homographs, name changes, and ill-defined specification or classification.

Synonyms, False Synonyms, and Homographs

Common terminology problems such as synonyms (different terms for the same thing), false synonyms (seemingly similar terms for different things), and homographs (the same term for different things) can assail both data and metadata. For example, the attribute named "region" in HGML sounds similar to "feature region" in GenBank, but they are in fact quite different things."Lys," "K," and "9," on the other hand, are all commonly used synonyms for data value codes in different databases for the same amino acid -- Lysine. Homographs are even more troublesome. For example, one laboratory uses the name "B1.1" for a probe that hybridizes to the MX1 locus, while another laboratory uses "B1.1" for a completely different probe that hybridizes to the APP locus.

Name Changes and Data

One of the problems with using names to connect information between databases is that names and their meaning may change over time. Some nomenclature systems such as [7] explicitly recognize that the name of any given entity may need to evolve to reflect scientific understanding of its composition or function. Such evolution includes not only simple renaming, but also differentiation (when what was originally thought to be a single entity turns out to be more than one) and consolidation (when what was originally thought to be two entities turn out to be the same) [8]. For example, the human locus associated with the disorder elliptocytosis was originally named EL. As evidence indicated that there were two loci, EL became EL1 and EL2. When specific gene products of those loci were identified, those names changed to PB41 (protein band 4.1) and SPTA (spectrin, alpha), respectively, and *EL became the allelle name [7].

Data and Metadata Specification and Classification

Information retrieval is difficult (at best) if attributes are not discrete, consistent, and well-defined. When multiple types of information are contained in a single attribute, it becomes more difficult to locate, process, and connect to related data. Data entity names are especially susceptible to information overload when people try to summarize several different characteristics in a single name, rather than as separate attributes [9].

Classifications need not follow a single hierarchy, but they must be consistent. If controlled vocabularies and classification schemes are not used for non-numeric attributes, locating such data becomes haphazard -- and linkage becomes impossible.

THE DATA THESAURUS: A NEW TOOL FOR TERMINOLOGY

At the same time as computers have posed new challenges, they also have enabled development of tools to address problems of data terminology and integration. The data thesaurus is one such software tool that deals with a number of the issues discussed in the preceding sections. Originally developed in conjunction with a prototype information system for material properties data [10; 11], the data thesaurus concept extends ideas from earlier metadata tools, including statistical codebooks, subject term thesauri, data dictionaries, and database schemas in order to:

- manage definitions and cross references for various types of metadata;
- reconcile diverse nomenclatures and classifications from multiple sources; and
- link information between different types of entities from federated databases.

General Architecture and Features

Conceptually, the data thesaurus is a logical layer that lies between the database management system (DBMS) and users or programs, as pictured in Exhibit 4. It includes a specialized database of metadata, special functions for indexing and access, and interfaces for people and software. It can be implemented using the database management system.

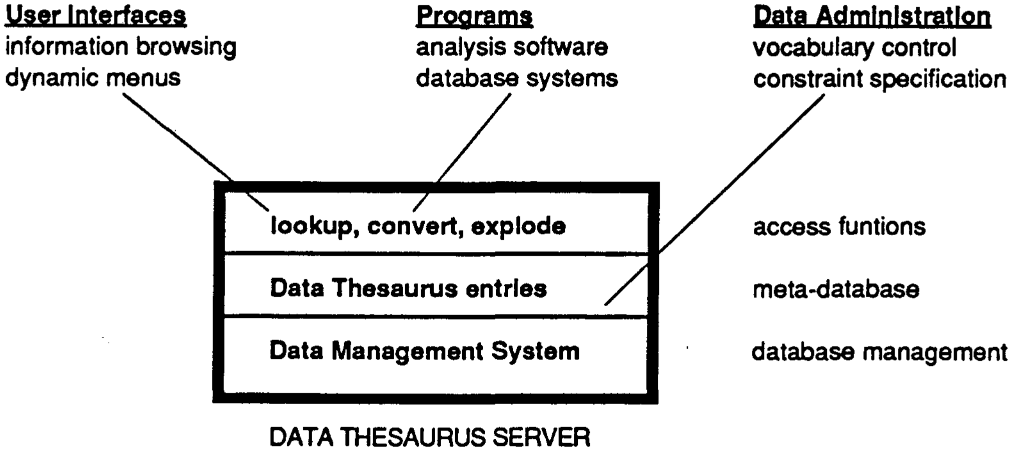


Exhibit 4: Modular Data Thesaurus Architecture supports people and computer programs

Thesaurus Paradigm and Components

A data thesaurus is a special type of text database that organizes metadata much as controlled vocabulary thesauri organize indexing terms for bibliographic information systems [12]. But whereas bibliographic thesauri contain a fixed set of attributes for one type of entity (subject terms), data thesauri include different components concerning multiple types of metadata entities and their relationships.

A data thesaurus is composed of individual *entries*, each of which represents an individual instance of a particular entity, attribute, or domain (set of comparable values). Each entry is in turn composed of discrete components that contain distinct types of metadata information. Exhibit 5, for example, shows a simplified hypothetical data thesaurus entry for a particular type of biotechnical entity -- namely a gene. Other entries might pertain to other types of entities (e.g., proteins), attributes (e.g., Location), or domain values (e.g., amino acid code values). In this example, components have been clustered into broad functional classes, as indicated at the left of the exhibit.

Function		Component	Component Value(s)		
Identity/Type		Identifier	2549685-7		
		Sequence Tag	...CAACGGTATAGGCTAACCTG ...		
		Type	gene		
Names/source(s)		Primary Name	SOD1		
		Prior Name (source)	IPO-A (mim)	SODS (hgml)	
		Used For	superoxide dismutase-1		
Internal links		Related Genes (RE)	SOD2		
		Alleles (NE)	SOD A*1	SOD A*2	
External links		MIM Reference(s)	14745		
		GenBank Sequence(s)	J02947	K00065	M13267
		HGML Locus Identifier	LM0147		
Book-keeping	Audit	Date Modified	6/28/89	2/1/89	3/4/88
	Trail	Modified By	jlmc	teh	bcd

Exhibit 5: Simplified Hypothetical Data Thesaurus Entry shows metadata components

Note that some components may be comprised of composite structures (e.g., Audit Trail contains two attributes). Some have multiple values (e.g., Alleles, Prior Name), and some may be constrained to specified domains (e.g., Sequence Tag contains only A,C,G, or T as values). Some components are pointers to other data thesaurus entries -- "RE" denotes Related Entries and "NE" denotes "Narrower Entries." As in object-oriented programming, broader class entities may contain generic information which all members of the class inherit, and classes may contain other classes recursively. Other components are identifiers (foreign keys) for data records in external databases.

Users and Uses

Data thesauri can be used in a variety of ways. Scientists may browse or query a thesaurus directly, or use it in conjunction with interface and data administration software to manage data and metadata integrity constraints for the DBMS; translate names into preferred terms or identifiers; "explode" a term to include broader, narrower, or related terms; convert between measurement units or coded value domains; and generate dynamic menus based on broader term/narrower term relationships.

The data thesaurus provides a bridge between diverse terminologies and databases. It does not impose a single nomenclature or classification standard. Instead, it provides a controlled environment within which multiple names, classification hierarchies, and value coding schemes can co-exist and evolve independently. It does not automatically resolve problems of terminology or data structure inconsistency, but it provides a framework for addressing such problems in a systematic way. Terminology managers can use the thesaurus to organize and document addition, deletion, and modification of entries over time (using attributes such as "Date Added," "Prior Term," etc.). Different individuals or groups may be responsible for disjoint sets of entities, attributes, or domains, so long as they do so within the over-all framework. Just as the metathesaurus currently being developed as part of the National Library of Medicine's Unified Medical Language System will unite controlled vocabularies from several bibliographic systems [13], the data thesaurus concept could be used to document and integrate work being done by various national and international standards efforts.

Implementation Issues

Experience with LBL's prototype material properties data thesaurus has demonstrated that implementation of this concept requires standard database capabilities plus special features for text data structures. LBL is currently exploring how different types of database software might meet these requirements. Commercial relational systems offer a mature, standard data model, query language (SQL), and program interfaces, but they are not well-suited for text and metadata [14]. Text management and retrieval systems offer much of the necessary text capabilities, but not a standard query language. Object-oriented database systems appear well-suited to many of our requirements, but few, if any have special features for text and the technology is untested.

SUMMARY CONCLUSIONS

The rapidly increasing amount, scope and diversity of biotechnical data made possible by computers has raised a host of terminology problems. Bioscientists urgently need better administrative and software mechanisms to develop and maintain controlled vocabularies and classification schemes necessary for effective data access, integration, and sharing. The data thesaurus is a new concept that addresses many of these issues. It builds on a familiar paradigm and previous metadata management tools to provide a systematic, extensible framework within which people and computer programs, standards organizations and individual research projects can:

- picture how different types of information relate to one another;
- reconcile nomenclature and classification systems from diverse, autonomous databases;
- support data and meta-data integrity constraints in a unified framework; and
- facilitate coordination between autonomous participating organizations.

ACKNOWLEDGEMENTS

This work was supported by the Director, Office of Energy Research, Health and Environmental Research, Human Genome Program, of the U.S. Department of Energy, under Contract DE-AC03-76SF00098.

REFERENCES

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

2 Ingerman PZ (1982) "Make the Most of Your Check Digits," *Dr. Dobb's Journal* 66:36

3 Gumm HP (1985) "A New Class of Check-Digit Methods for Arbitrary Number Systems," *IEEE Transactions on Information Theory* IT-31:1 pp. 102-105.

4 McCarthy JL (1984) "Scientific Information = Data + Metadata," In Solomon H (Ed.), *Proceedings of the Workshop on Data Base Management, Monterey, CA* , pp. 79-124.

5 ANSI Acredited Standards Committee X3L8 (1990) "Coordination of Data Elements,"

6 Board of Regents NLM (1986) "Report of the Board of Regents, National Library of Medicine, NLM Long Range Plan Report of Panel 3: Obtaining Factual Information from Databases," (U.S. Department of Health and Human Services).

7 ISGN (1987)*Guidelines for Human Gene Nomenclature: An International System for Human Gene Nomenclature (ISGN, 1987),* (S. Karger)

8 Fleck L. *Genesis and Development of a Scientific Fact* . (Chicago: U. Chicago).

9 Newton JJ (1987) "Guide on Data Entity Naming Conventions," *NBS Special Publication 500-149* (U.S. Department of Commerce, National Bureau of Standards).

10 McCarthy JL (1987) "Information Systems Design for Material Properties Data," In Glazman JS, Rumble JR Jr. (Ed.), *Proceedings of the Computerization and Networking of Materials Data Bases, Philadelphia* , pp. 135-150.

11 McCarthy JL (1988) "The Automated Data Thesaurus: A New Tool for Scientific Information," In CODATA (Ed.), *Proceedings of the 11th International CODATA*

Conference, Karlsruhe, Germany

12 American National Standards Institute. Subcommittee 25 on Thesaurus Rules and Conventions. (1980) "American National Standard Guidelines for Thesaurus Structure, Construction, and Use," *ANSI Z39.19-1980, revision of ANSI Z39.19-1974*

13 Humphreys BL, Lindberg DAB (1989) "Building the Unified Medical Language System," *Proc Symp on Computer Applications in Medical Care, Washington, D.C.*

14 Lynch CA (1988) "Developments in Database Management System Technology and Their Impact on Information Retrieval," *Proc. ASIS Annual Meeting,* , pp. 190-194.

LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
INFORMATION RESOURCES DEPARTMENT
1 CYCLOTRON ROAD
BERKELEY, CALIFORNIA 94720