# APPENDIX A – Standards Most Relevant for NCI Use

## 1. COMMON DEMOGRAPHIC/INFORMATION PROCESSING AND CODE SETS

This section addresses data standards for commonly collected information, including demographic characteristics, locational information, and the like. The general information standards might be of interest across federal agencies and of use to organizations conducting surveys and statistical analysis.

## 1.1 Address

### 1.1.1 National Institute of Standards and Technology (NIST) Federal Information Processing Standards (FIPS)

<http://www.itl.nist.gov/fipspubs/>

***Sponsor***

The National Institute of Standards and Technology (NIST)

***Description***

Under Section 513 of the Information Technology Management Reform Act of 1996 and the Computer Security Act of 1987, Public Law 104-106, NIST develops standards, guidelines, and associated methods and techniques for federal computer systems. NIST issues these standards and guidelines as Federal Information Processing Standards (FIPS) for use government-wide. FIPS codes are a standardized set of numeric or alphabetic codes issued to ensure uniform identification of geographic entities through all federal government agencies.

FIPS are developed only when there are no existing voluntary standards to address federal requirements for the interoperability of different systems, for the portability of data and software, and for computer security. NIST collaborates with national and international standards committees, users, industry groups, consortia, and research and trade organizations, to develop standards.

***Usage***

The FIPS codes are applicable to systems requiring the exchange of data among users and internal data systems where such use contributes to operational benefits, efficiency, and economy.

***Applicability to NCI***

Applicable FIPS publications are described below. Some of these standards are redundant with others. During the harmonization process, NCI should review various external standards, along with data elements already registered in the caDSR, to decide what would be the most useful to register.

#### *1.1.1.1* FIPS 5-2*, Codes for Identification of the States, District of Columbia and Outlaying Areas of the United States, and Associated areas.*

CODES FOR THE IDENTIFICATION OF THE STATES, THE DISTRICT OF COLUMBIA AND THE OUTLYING AREAS OF THE UNITED STATES, AND ASSOCIATED AREAS, 1987 May 28. This standard provides a set of 2-digit numeric codes and a set of 2-letter alphabetic codes for representing the 50 states, the District of Columbia and the outlying areas of the U.S., and associated areas such as the Federated States of Micronesia and Marshall Islands, and the trust territory of Palau.

#### *1.1.1.2 FIPS 10-4, Countries, Dependencies, Areas of Special Sovereignty, and their Principal Administrative Divisions.*

COUNTRIES, DEPENDENCIES, AREAS OF SPECIAL SOVEREIGNTY, AND THEIR PRINCIPAL ADMINISTRATIVE DIVISIONS, 1995 April (reflects technical changes through May 6, 1993). This standard provides a list of the basic geopolitical entities in the world, together with the principal administrative divisions that comprise each entity. Each basic geopolitical entity is represented by a 2-character, alphabetic country code. Each principal administrative division is identified by a 4-character code consisting of the 2-character country code followed by a 2-character administrative division code. This data element and representation standard is made available for the interchange of information among Federal departments and independent agencies. It is intended for use in activities associated with the mission of the Department of State and national defense programs. It may also be used for Federal interchanges of information with the nonfederal sector including industry, State, local, and other governments, and the public at large.

### 1.1.2 International Organization for Standardization (ISO)

<http://www.iso.org/>

***Sponsor***

The International Organization for Standardization (ISO).

***Description***

ISO is a network of national standards institutes in 148 countries. As a non-governmental organization, ISO has no legal authority to enforce their implementation. A certain percentage of ISO standards - such as those concerned with health, safety or the environment - have been adopted in countries as part of their regulatory framework, or are referred to in legislation for which they serve as the technical basis. ISO develops standards for which there is a market requirement. Experts conduct the work in industrial, technical, and business sectors that have asked for the standards and subsequently used them. Others with relevant knowledge, such as representatives of government agencies, consumer organizations, academia and testing laboratories, may join these experts. Although ISO standards are voluntary, they are developed in response to market demand and are based on consensus among the interested parties, thus ensuring widespread applicability of the standards.

### *Usage*

ISO is involved with a wide range of standards and they are in use by a wide variety of organizations worldwide. This report section has focused on the narrow subset of commonly used ISO data standards. In some cases, ISP standards address topics for which there is a competing national standard (for example, date or country names), and organizations need to evaluate which standard best serves their purposes. NCI would not be allowed to redistribute the ISO standards free of charge. However, registration of the data elements and value lists associated with the standards is not a problem as they are already available from other registries. NCI would not be allowed to redistribute the ISO standards free of charge. However, registration of the data elements and value lists associated with the standards is not a problem as they are already available from other registries.

### *Curation*

The ISO values selected for use at NCI can be curated as lists in EVS and/or the caDSR as value domains.

### *NCI Role*

Organizations have a range of opportunities for taking part in ISO's work, or in contributing to the development of standards through the ISO member in their country. Individuals may be selected by member institutes to serve on national delegations participating in ISO technical committees, or may provide their input during the process of developing a national consensus for presentation by the delegation. International organizations and associations, both nongovernmental and representing industry sectors, can apply for liaison status to a technical committee. They do not vote, but they can participate in the debates and the development of consensus. However, general ISO data standards are not within NCI's sphere of interest, so it is not likely that NCI will choose to actively participate in the development of these standards.

### *Point of Contact*

ISO Central Secretariat
International Organization for Standardization (ISO)
1, rue de Varembé, Case postale 56
CH-1211 Geneva 20, Switzerland
Telephone: +41 22 749 01 11

The ISO member body organization for the United States is:

American National Standards Institute (ANSI)
1819 L Street, NW.
Washington, DC 20036
(202) 293-8020
E-mail: info@ansi.org

#### *1.1.2.1 ISO 3166-1, Codes for the representation of names of countries and their subdivisions - Part 1: Country codes.*

This standard contains a 2-letter code, which is recommended as the general purpose code, a 3-letter code which has better mnemonic properties, and a numeric-3 code, which can be useful if script independence of the codes is important.

***Applicability to NCI***

This standard would be applicable to systems requiring the exchange of county codes.

***License***

The short country names from ISO 3166-1 and the alpha-2 codes are made available by ISO at no charge for internal use and non-commercial purposes. The incorporation of the complete ISO 3166-1 standard document in commercial products may be subject to a charge. The lists extracted from ISO 3166-1 contain all short country names and alpha-2 code elements officially published by ISO and are updated whenever a change of country name and/or alpha-2 code element is made in ISO 3166-1. Country names and code elements can be downloaded from the ISO Web site in html, text, Extensible Markup Language (XML), and Microsoft Access formats.

#### *1.1.2.2 ISO 11180:1993, Postal addressing*

This standard specifies the maximum dimensions of the postal address and its locations on forms complying with ISO 8439 and is designed to standardize its presentation and structure. Annexes A and B of the standard give elements of the addressee's address and examples of addresses.

***License***

The Postal Addressing standard can be purchased from the ISO web site.

***Applicability to NCI***

This standard should be considered when developing a postal addressing standard but may be superseded by broader standards, such as the one developed by the Universal Postal Union.

### 1.1.3 Universal Postal Union Standards

<http://www.upu.int/>

***Sponsor***

Universal Postal Union (UPU)

### *Description*

The UPU is the primary forum for cooperation between postal services and strives to ensure a universal network of up-to-date products and services. In this way, the organization fulfils an advisory, mediating and liaison role, and renders technical assistance where needed. It sets the rules for international mail exchanges and makes recommendations to stimulate growth in mail volumes and to improve the quality of service for customers.

### *Usage*

Standards are important prerequisites for effective postal operations and for interconnecting the global network. The UPU's Standards Board develops and maintains a growing number of standards to improve the exchange of postal-related information between posts and promotes the compatibility of UPU and international postal initiatives. It works closely with posts, customers, suppliers and other partners, including various international organizations. The Standards Board ensures that coherent standards are developed in areas such as electronic data interchange (EDI), mail encoding, postal forms, and meters.

### *Applicability to NCI*

This standard would be applicable to systems requiring the exchange of mailing address data among users and internal data systems where such use contributes to operational benefits, efficiency, and economy. This standard includes addressing specifications and state codes for countries participating in the UPU. Because NCI supports a variety of activities in many countries, use of this standard will be necessary to ensure that mailed information is properly addressed.

### *Curation*

UPU maintains a set of code lists available on the UPU Web site. They could be added to the caDSR as value domains if deemed useful.

### *NCI Role*

The Standards Board consists of postal administrations from countries that are committed to the development of standards within the postal community. Members of the Standards Board are experts appointed on the basis of their qualifications in international postal operations and information technologies. Congress appoints the chairing country of the Standards Board. Any UPU member administration may become a full member of the Standards Board. There are a number of permanent working groups that have been established under the responsibility of the Standards Board. These groups have been created to provide permanent and consistent coverage of particular working areas in standardization. Any UPU member administration may become a member of one or more of the Standards Board's working groups.

### *Point of Contact*

Universal Postal Union
International Bureau
Case postale 13
3000 BERNE 15
SWITZERLAND

Telephone: +41 31 350 31 11
info@upu.int

## 1.2 Language

### 1.2.1 ISO 639, Codes for representation of language

ISO 639 Codes for the representation of names of languages--Part 1: Alpha-2 code and Part 2: Alpha-3 code provide language code elements comprising 2-letter and 3-letter language identifiers for the representation of names of languages. The language identifiers were devised for use in terminology, lexicography, and linguistics, but may be adopted for any application requiring the expression of language in coded form, especially in computerized systems.

***Applicability to NCI***

This standard is applicable to systems requiring the identification and exchange of language information, such as the language used in a document or dataset, the primary language spoken by a study participant, or the language used on a form.

## 1.3 Race and Ethnicity

### 1.3.1 Office of Management and Budget (OMB) Standards for the Classification of Federal Data on Race and Ethnicity

<http://www.whitehouse.gov/omb/fedreg/ombdir15.html>

***Sponsor***

The Office of Management and Budget (OMB) sponsors the Federal Race and Ethnicity standards in order to ensure that a consistent standard is used for recording race and ethnicity across the federal government. An interagency committee developed the actual standard codes:

"OMB announced in July 1993 that it would undertake a comprehensive review of the current categories for data on race and ethnicity. This review has been conducted over the last four years in collaboration with the Interagency Committee for the Review of the Racial and Ethnic Standards, which OMB established in March 1994 to facilitate the participation of Federal agencies in the review.

The members of the Interagency Committee, from more than 30 agencies, represent the many and diverse Federal needs for data on race and ethnicity, including statutory requirements for such data."[1]

The code set was updated in 1995 and 1997. It is likely that future revisions of the code set will be handled in a similar way. The standard was made a part of OMB's Statistical Policy Directive No. 15, Race and Ethnic Standards for Federal Statistics and Administrative Reporting, which required agencies to use the standard in all reporting.

[1] OMB website: < http://www.whitehouse.gov/omb/fedreg/directive_15.html>

### *Description*

In the United States, federal standards for classifying data on race determine the categories used by federal agencies and exert a strong influence on categorization by state and local agencies and private sector organizations.  The basic concepts are defined as follows for use in federal statistics:

**Race**: The federal standards do not conceptually define race, and they recognize the absence of an anthropological or scientific basis for racial classification.  Instead, the federal standards acknowledge that race is a social-political construct in which an individual's own identification with one or more race categories is preferred to observer identification.  The standards use a variety of features to define five minimum race categories.  Among these features is descent from "the original peoples" of a specified region or nation.  The minimum race categories are American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, and White.  The federal standards stipulate that race data need not be limited to the five minimum categories, but any expansion must be collapsible to those categories.

**Ethnicity**: The federal standards do not conceptually define ethnicity, and they recognize the absence of an anthropological or scientific basis for ethnicity classification.  Instead, the federal standards acknowledge that ethnicity is a social-political construct in which an individual's own identification with a particular ethnicity is preferred to observer identification.  The standards specify two minimum ethnicity categories: Hispanic or Latino, and Not Hispanic or Latino.  The standards define a Hispanic or Latino as a person of "Mexican, Puerto Rican, Cuban, South or Central America, or other Spanish culture or origin, regardless of race."  The standards stipulate that ethnicity data needs to be limited to the two minimum categories, but any expansion must be collapsible to those categories.  In addition, the standards stipulate that an individual can be Hispanic or Latino or can be Not Hispanic or Latino, but not both.

### *Usage*

According to OMB, "The standards are used not only in the decennial census (which provides the "denominator" for many measures), but also in household surveys, on administrative forms (e.g., school registration and mortgage lending applications), and in medical and other research."[2] These race and ethnicity categories have an extremely wide currency.  This situation can be seen in the fact that the OMB sponsored code set is also published by CDC’s Pharmaceutical and Healthcare Industry News (PHIN), and are offered as externally defined value sets by HL7.

### *Applicability to NCI*

The OMB sponsored race and ethnicity value sets should be used when this data is collected in the context of an NCI sponsored effort.  Some cancer research may require additional race and ethnic categories and these can be used in information collection and analysis.  All categories used need to be mapped to these standard OMB categories.

### *Curation*

Value domains should be created in the caDSR to incorporate these standard code values.

### *NCI Role*

[2] Ibid

The race and ethnicity categories are maintained though periodic federal government initiatives - managed by OMB - and are not created by a standards organization. As such, there are no defined structures for allowing input from parties outside of designated interagency committee. If there is an interest in having NCI involvement in revision to these code sets, this matter should be addressed by seeking membership in the relevant interagency committee.

## 1.4 Occupation Classification

### 1.4.1 Bureau of Labor Statistics, Standard Occupational Classification (SOC) System

<http://www.bls.gov/soc/>

***Sponsor***

The Bureau of Labor Statistics (BLS) is the primary developer of the SOC. However, a broadly based interagency committee formally sponsors the development of the data.
"In 1994, the Office of Management and Budget established a Standard Occupational Classification Revision Policy Committee (SOC Committee) to develop a unified classification structure that would meet the needs of the 21st century. The Committee was chaired by the Bureau of Labor Statistics and the Bureau of Census, with representatives from the Bureau of Labor Statistics, the Bureau of Census, the Employment and Training Administration (Department of Labor), the Office of Personnel Management, the Defense Manpower Data Center, and ex officio, the National Science Foundation, the National Occupational Information Coordinating Committee, and the Office of Management and Budget."[3]

***Description***

The occupational classifications are intended to describe the rules that are followed to classify workers based on occupational definition and work activity. The BLS documentation notes that: "The Standard Occupational Classification (SOC) System was developed in response to a growing need for a universal occupational classification system. Such a classification system would allow government agencies and private industry to produce comparable data. Users of occupational data include government program managers, industrial and labor relations practitioners, students considering career training, job seekers, vocational training schools, and employers wishing to set salary scales or locate a new plant. It will be used by all federal agencies collecting occupational data, providing a means to compare occupational data across agencies. It is designed to cover all occupations in which work is performed for pay or profit, reflecting the current occupational structure in the United States." [4]

The descriptive material goes on to state that:

"The SOC classifies workers at four levels of aggregation: 1) major group; 2) minor group; 3) broad occupation; and 4) detailed occupation. All occupations are clustered into one of 23 major groups. Within these major groups are 96 minor groups, 449 broad occupations, and 821 detailed occupations. Occupations with similar skills or work activities are grouped at each of the four levels of hierarchy to facilitate comparisons. For example, "Life, Physical and Social Science

[3] iii, Revising the Standard Occupational Classification System, Report 929, June 1999, Bureau of Labor Statistics

[4] < http://www.bls.gov/soc/socguide.htm>

Occupations" (19-0000) is divided into four minor groups, "Life Scientists" (19-1000), "Physical Scientists" (19-2000), "Social Scientists and Related Workers" (19-3000), and "Life, Physical and Social Science Technicians" (19-4000). Life Scientists contains broad occupations such as "Agriculture and Food Scientists" (19-1010), and "Biological Scientists" (19-1020). The broad occupation Biological Scientists includes detailed occupations such as "Biochemists and Biophysicists" (19-1021), and "Microbiologists" (19-1022)." [5]

***Usage***

The BLS Web site notes that the SOC is used by all government agencies that collect and publish occupational data. Within the healthcare arena, there has not been wide or consistent attention paid to occupational data. To the extent that this data was collected, the tendency was to use custom coding systems. However, the intent for HL7 Version 3 messaging is to use the Standard Occupational Classifications as the code system for occupational and workplace categories.

***Applicability to NCI***

The SOC Committee-sponsored occupational classifications should be used when this data is collected in the context of an NCI sponsored effort. It should be noted that SNOMED also includes occupational codes that may need to be coordinated with the SOC list.

***Curation***

The code set should be added to the EVS.

***NCI Role***

The standard occupational categories, maintained though periodic federal government initiatives (managed by the BLS) are not created by a standards organization. As such, there are no defined structures for allowing input from parties outside of designated interagency committees. NCI can provide input to the relevant interagency committee.

## 1.5 Vital Statistics

### 1.5.1 CDC National Center for Health Statistics (NCHS)

<http://www.cdc.gov/nchs>

***Sponsor***

The National Center for Health Statistics (NCHS), while formerly an independent agency within the Department of Health and Human Services (HHS), currently is a center within the Centers for Disease Control (CDC). The organization compiles statistical information to guide actions and policies to improve US health status. The NCHS Website states: "Working with partners throughout the health community, we use a variety of approaches to efficiently obtain information from the sources most able to provide information. We collect data from birth and death records, medical records, interview surveys, and through direct physical exams and laboratory testing. NCHS is a key element of our national public health infrastructure, providing important surveillance information that helps identify and address critical health problems."

[5] Ibid.

### *Description*

The NCHS is an organization that collects and manages data. The generation of standards is a necessary but secondary aspect of this primary activity. Data collection activities and NCHS standards are best discussed under two headings: collection of health-related data though survey activity, and collection of vital statistics data - births and deaths - through the receipt of periodic reports from vital statistics departments on the state level.

### *Surveys*

The following lists the surveys currently carried out by the NCHS:

- National Health and Nutrition Examination Survey (NHANES).
- National Health Care Survey (NHCS).
    - National Ambulatory Medical Care Survey (NAMCS).
    - National Hospital Ambulatory Care Survey (NHAMCS).
    - National Survey of Ambulatory Surgery (NSAS).
    - National Hospital Discharge Survey (NHHS).
    - National Nursing Home Survey (NHHCS).
    - National Home and Hospice Care Survey (NHHCS).
    - National Employer Health Insurance Survey (NEHIS).
    - National Health Provider Inventory (NHPI).
- National Health Interview Survey (NHIS).
- National Immunization Survey (NIS).
- National Survey of Family Growth (NSFG).
- State and Local Area Integrated Telephone Survey (SLAITS).

Collecting this data occurs periodically and uses various instruments. Data is collected though direct patient interviews, through telephone interviews, through extraction from medical and provider records, and through direct examination at a mobile examination site. NCHS provides documentation of the data that is collected by providing documentation on the questions that are asked, the layout of the data files made available to researchers, and through specifying the criteria used to constrain the data collected for a particular item. In some cases, data items are constrained through specification that responses must be drawn from a particular code set - in these cases the valid codes are defined directly within the survey and data output documentation. There does not seem to be a central representation of the attributes collected across the body of surveys.

### *Vital Statistics*

The NCHS Website notes, "The National Vital Statistics System is responsible for the Nation’s official vital statistics. These vital statistics are provided through State-operated registration systems. The registration of vital events - births, deaths, marriages, divorces, and fetal deaths - is a state function.

Standard forms for the collection of the data and model procedures for the uniform registration of the events are developed and recommended for state use through cooperative activities of the States and NCHS. The process for implementing revisions for the birth and death certificates and the fetal death report is now underway. Current drafts of the revised certificates and accompanying technical information are available. NCHS shares the costs incurred by the States in providing vital statistics data for national use. NCHS also produces training and instructional material."

Vital statistics data is collected in the following areas:

- Birth Data
- Mortality Data
- Marriages and Divorces. (Collection of detailed marriage and divorce data was suspended in January 1996.)
- Fetal Death

For each of these areas, NCHS has defined a positional file format that defines the valid fields (data elements) for each file and provides value domains for coded elements. Data is received periodically in files whose layouts correspond to the file formats. Where a single element is used in multiple files, e.g., race and ethnicity, the same definition is used across those files. However, as with the surveys, there does not appear to be a formally documented model or standard that underlies the file layouts.

### *Usage*

The surveys are carried out regularly by the NCHS, and researchers and policy analysis use the results of the survey for wide-ranging purposes. In a similar manner, birth and death information is received from states to support national vital statistics reporting, and is used to support uses ranging from evaluation of different causes of death through reviewing Social Security applications.

While the data collected by NCHS is widely used, the standards behind that data are not used beyond the fact of their implicit appearance within the data itself.

### *Applicability to NCI*

The NCHS standards for vital statistics data are likely to have some relevance to NCI researchers in those cases in which birth information, birth abnormalities, and death information is relevant to research activities. In these cases, it will be valuable to have access to the NCHS's standards for collecting this data.

### *Curation*

Given the potential relevance to NCI, there would be some value in including the NCHS formats for birth and death information within the NCI repositories. This task would be complicated by the fact that NCHS provides the standard formats as fixed format record layouts, and embeds code values within the entries for the relevant fields. Reformatting of the data would be needed to make it conform to the requirements of a metadata registry.

### *NCI Role*

Given that NCHS is a government department rather than a standards developer, there is limited scope for direct NCI involvement in its work. To the extent that there is NCI interest in influencing the standards that underlay NCHS data collection activity, this is probably best addressed through the HHS Data Council and CDC's Public Health Information Network (PHIN).

## 1.6 Measurement

### 1.6.1 ISO 31, Quantities and units

This standard specifies individual standards dealing with quantities in space and time, periodic phenomena, mechanics, heat, electricity and magnetism, electromagnetic radiation, chemistry, molecular physics, nuclear physics, etc., especially the International System of Units, SI, including recommendations for printing symbols and numbers.

***Applicability to NCI***

This standard is applicable to systems that use and report quantities and units of measure and should be one of the standards considered when creating a list of units of measure to meet the broad needs of NCI applications. While the standard does not specifically include units of measure for all medical or laboratory applications, it does include space and time, chemistry, and nuclear physics.

### 1.6.2 HL7 Units (Versions 2.X+)

HL7 has adopted codes for Units (Versions 2.X +), derived from the ISO 2955-83 standard (withdrawn by ISO in 2001) and ANSI X3.50. This standard will be used to define common units of measure, such as Celsius or mg/ml, that are intended to be combined with a numeric value to accurately express a result.

***Applicability to NCI***

This standard is applicable to systems that must use HL7 messages for transmitting information. The HL7 list should be a candidate for adoption as the standard list for use in the caDSR value domain specifications. The content of the list is currently not fully specified for HL7 Version 3.0.

## 1.7 Information Processing

### 1.7.1 FIPS 4-2, REPRESENTATION OF CALENDAR DATE FOR INFORMATION INTERCHANGE, 1998 November 15.

This standard provides a means of representing calendar date to facilitate interchange of data among information systems. This standard adopts American National Standard Institute (ANSI) X3.30-1997: Representation of Date for Information Interchange (revision of ANSI X3.30-1985 (R1991). This standard was issued to preclude any confusion about the use of date format standards within the federal government, and to carry out the objectives of the President's Council on Year 2000 Conversion. FIPS 4-2 supersedes FIPS PUB 4-1, dated January 27, 1988, and updates the standard for representing calendar date and implements the Federal Government's commitment to use 4-digit year elements (e.g., 1999, 2000, etc.) in its information technology systems.

### 1.7.2 ISO 8601, Numeric representation of dates and time

This standard offers representations for the following: date; time of the day; coordinated universal time (UTC); local time with offset to UTC; date and time; time intervals; and recurring time intervals. ISO 8601 is used in computer programs, logbooks, contest entries, magazine reports, Web pages, e-mail, statistics, forms of all kinds, administrations and businesses, in customs and transportation, in e-commerce and academia, and in all types of international activity.

***Usage***

Versions of the ISO 8601 have been adopted by HL7 as well as other standards bodies. The FIPS standard is frequently adopted by federal agencies. These standards generally apply to data exchange formats.

***Applicability to NCI***

NCI has many data elements related to date and time. It is difficult to apply date standards to data storage formats. It may be more useful to standardize the data exchange formats.

***Curation***

The NCI Context Administrators will review the Date and Time standards as part of the harmonization process. Selected date and time data elements will be registered in the caDSR as part of the NCI context.

## 2. HEALTH-RELATED VOCABULARY/CODING STANDARDS

This section addresses a number of vocabulary data standards designed to foster standardization in collection, exchange, and storage of health-related information. These standards mostly consist of lists of concept names or codes and related definitions.

## 2.1 Basic Biology Vocabularies

### 2.1.1 International Union of Biochemistry and Molecular Biology (IUBMB) and the International Union of Pure and Applied Chemistry (IUPAC)

<http://www.iubmb.unibe.ch/>
<http://www.iupac.org/index_to.html>

***Sponsor***

The IUBMB is an international organization of biochemists and molecular biologists, while the IUPAC is an international union of chemists. They jointly developed nomenclature recommendations for many biochemical and chemical substances and processes that are relevant to biological research.

***Description***

IUPAC-IUBMB Joint Commission on Biochemical Nomenclature (JCBN) and the Nomenclature Committee of the IUBMB (NC-IUBMB) develop and publish nomenclature recommendations for the following categories of biochemicals:

| **Recommendation** | **URL** |
|---|---|
| Amino Acids and Peptides | http://www.chem.qmul.ac.uk/iupac/AminoAcid/ |
| Biochemical thermodynamics | http://www.chem.qmul.ac.uk/iubmb/thermod/ |
| Branched nucleic acids | http://www.chem.qmul.ac.uk/iubmb/misc/bran.html |

| Recommendation | URL |
|---|---|
| Carbohydrates | http://www.chem.qmul.ac.uk/iupac/2carb/ |
| Carotenoids | http://www.chem.qmul.ac.uk/iupac/carot/ |
| Corrinoids (vitamin $B_{12}$) | http://www.chem.qmul.ac.uk/iupac/misc/B12.html |
| Cyclitols | http://www.chem.qmul.ac.uk/iupac/cyclitol/ |
| Electron transport proteins | http://www.chem.qmul.ac.uk/iubmb/etp/ |
| Enzyme kinetics | http://www.chem.qmul.ac.uk/iubmb/kinetics/ |
| Enzyme nomenclature | http://www.chem.qmul.ac.uk/iubmb/enzyme/ |
| EC 1 Oxidoreductases | http://www.chem.qmul.ac.uk/iubmb/enzyme/EC1/ |
| EC 2 Transferases | http://www.chem.qmul.ac.uk/iubmb/enzyme/EC2/ |
| EC 3 Hydrolases | http://www.chem.qmul.ac.uk/iubmb/enzyme/EC3/ |
| EC 4 Lyases | http://www.chem.qmul.ac.uk/iubmb/enzyme/EC4/ |
| EC 5 Isomerases | http://www.chem.qmul.ac.uk/iubmb/enzyme/EC5/ |
| EC 6 Ligases | http://www.chem.qmul.ac.uk/iubmb/enzyme/EC6/ |
| Folic acid | http://www.chem.qmul.ac.uk/iupac/misc/folic.html |
| Glycolipids | http://www.chem.qmul.ac.uk/iupac/misc/glylp.html |
| Glycoproteins | http://www.chem.qmul.ac.uk/iupac/misc/glycp.html |
| *myo*-Inositol numbering | http://www.chem.qmul.ac.uk/iupac/cyclitol/myo.html |
| Lignan Nomenclature | http://www.chem.qmul.ac.uk/iupac/lignan/ |
| Lipid Nomenclature | http://www.chem.qmul.ac.uk/iupac/lipid/ |
| Multienzymes | http://www.chem.qmul.ac.uk/iubmb/misc/menz.html |
| Multiple forms of enzymes | http://www.chem.qmul.ac.uk/iubmb/misc/isoen.html |
| Nucleic acid constituents | http://www.chem.qmul.ac.uk/iupac/misc/naabb.html |
| Nucleic acid sequence (incompletely specified bases) | http://www.chem.qmul.ac.uk/iubmb/misc/naseq.html |
| Peptide hormones | http://www.chem.qmul.ac.uk/iubmb/misc/phorm.html |
| Phosphorus containing compounds | http://www.chem.qmul.ac.uk/iupac/misc/phospho.html |
| Polymerized amino acids | http://www.chem.qmul.ac.uk/iupac/misc/polypep.html |
| Polypeptide conformation | http://www.chem.qmul.ac.uk/iupac/misc/ppep1.html |
| Polynucleotide conformation | http://www.chem.qmul.ac.uk/iupac/misc/pnuc1.html |
| Polysaccharide conformation | http://www.chem.qmul.ac.uk/iupac/misc/psac.html |
| Prenol nomenclature | http://www.chem.qmul.ac.uk/iupac/misc/prenol.html |
| Pyridoxal (vitamin $B_6$) | http://www.chem.qmul.ac.uk/iupac/misc/B6.html |
| Quinones with an Isoprenoid Chain | http://www.chem.qmul.ac.uk/iupac/misc/quinone.html |
| Retinoids | http://www.chem.qmul.ac.uk/iupac/misc/ret.html |
| Steroids | http://www.chem.qmul.ac.uk/iupac/steroid/ |
| Tetrapyrroles | http://www.chem.qmul.ac.uk/iupac/tetrapyrrole/ |
| Tocopherols (vitamin E) | http://www.chem.qmul.ac.uk/iupac/misc/toc.html |
| Translation Factors | http://www.chem.qmul.ac.uk/iubmb/misc/trans.html |
| Vitamin D | http://www.chem.qmul.ac.uk/iupac/misc/D.html |

These naming cover a wide range of bio- and other chemicals, as well as biochemical processes.

***License***

The IUBMB-IUPAC recommendations are open and published for anyone to use.

***Usage***

The IUBMB-IUPAC recommendations should be used by scientists to name newly discovered compounds or substances. For historical reasons some biochemicals and biochemical reactions in publications and databases are often referred to by their non-standard names, and because some of the IUBMB-IUPAC recommendations (e.g. enzyme nomenclature from the Enzyme Commission) are still being modified. In such situations, the recommended nomenclature should be added as aliases or synonyms where possible so that the same entity referred in different sources can be identified by their standard names. For example, the Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic pathways were developed using the recommended enzyme nomenclature and biochemical reactions.

***Applicability to NCI***

IUBMB nomenclature for biochemistry and molecular biology, particularly enzymes, may be particularly suited to NCI research. As the IUBMB is a long-standing standards organization, this is a credible standard to be used in NCI or by its partner organizations.

The Enzyme Commission (EC) organized by the precursor of the IUBMB created an enzyme classification scheme based on the reactions catalyzed by enzymes. The scheme is four levels deep; the first three levels are classes, and it includes a four-field numeric entry reflecting the tree placement (the EC number) and a name; the fourth level can be used as a name. When a new enzyme is identified, they assign the name and an EC number. The EC numbers and names are not unique. [Any number of the various protein kinases that are involved in signal transduction will have the same EC number and name because they essentially catalyze the same reaction.]

It should also be noted that the Consolidated Health Informatics (CHI) Initiative has recommended the Environmental Protection Agency (EPA) Substance Registry System (SRS) as a source for codes for chemicals of importance to health care outside of medications.

***Curation***

The NCI has added EC numbers and names to various enzymes as synonyms in the Metathesaurus. The enzyme hierarchy in the NCI Thesaurus is modeled after the EC class hierarchy.

To fully include the IUBMB enzyme information in the NCI Thesaurus, a dedicated EC property would need to be added. An alternative is to include it as a standalone vocabulary.

***NCI role***

None at this time.

## 2.2 Clinical Vocabularies and Code Sets

### 2.2.1 Common Terminology Criteria for Adverse Events (CTCAE) v3.0

***Sponsor***

The standard was developed by the Cancer Therapy Evaluation Program (CTEP), National Cancer Institute (NCI).

***Description***

The Common Toxicity Criteria (CTC) was developed in 1982 for use in adverse drug experience reporting, study adverse event (AE) summaries, Investigational New Drug (IND) reports to the Food and Drug Administration (FDA), and publications. The most recent version, CTCAE v3.0 (Common Terminology Criteria for Adverse Events version 3.0) represents the first comprehensive, multimodality grading system for reporting the acute and late effects of cancer treatment. The new CTC requires changes in the application of AE criteria including new guidelines regarding late effects, surgical and pediatric effects, multimodality issues, and for reporting the duration of an effect. It builds on the strengths of previous systems, represents a considerable effort among hundreds of participants, and signifies an international collaboration and consensus of the oncology research community.

The CTCAE v3.0 and its associated grading criteria are very specific with the intent to ensure that, wherever possible, each AE represents a single clearly definable clinical entity. In most instances, the CTCAE v3.0 provides an AE term and grade that more precisely describes the event, or provides a term for AEs heretofore unclassifiable. The compilation of AEs used to describe an incident provides a more complete characterization of the events that occur; they do not necessarily indicate more toxic interventions. The goal of the CTCAE v3.0 is to facilitate and improve the descriptions of AEs that do occur.

***Usage***

The NCI CTC v2.0 has become the worldwide standard dictionary for reporting acute AEs in cancer clinical trials and has been translated into several languages.

***Applicability to NCI***

CTCAE v3.0 includes Adverse Events applicable to all oncology clinical trials regardless of chronicity or modality. It is applicable to all clinical trials work at the NCI.

***Curation***

NCI plans to include CTCAE v3.0 in EVS.

***NCI Role***

NCI's CTEP program is the sponsor of the standard.

### 2.2.2 International Classification of Diseases for Oncology (ICD-O-3), Third Edition

<http://w3.whosea.org/rdoc/rdoc/publication.asp?bkid=221>

***Sponsor***

The World Health Organization (WHO) is the United Nations specialized agency for health. WHO's objective, as set out in its constitution, is the attainment by all peoples of the highest possible level of health. Health, defined in WHO's constitution as a state of complete physical, mental and social well-being, is not merely the absence of disease or infirmity. WHO publishes the *International Classification Of Diseases For Oncology* (ICD-O).

***Description***

ICD-O has been used for nearly 25 years as a standard tool for coding diagnoses of neoplasms in tumor and cancer registrars and in pathology laboratories throughout the world. ICD-O is a dual classification with coding systems for both topography and morphology. The 10-digit code describes where the tumor arose (a 4-character topography code for the primary site), what the tumor is (a 4-digit histology code for the cell type), how it behaves (a 1-digit code for malignant, benign, and so forth), and how aggressive it is (a 1-digit code for differentiation or grade).

The first edition of ICD-O was published in 1976, and a substantial revision - primarily of the topography codes – was published in 1990. Substantial changes have occurred over the past decade in techniques for diagnosing neoplasms. As a result, pathologists have been able to provide much more specific information about certain cancers. In particular, cytogenetic techniques have added considerably to the body of knowledge about malignant lymphomas and leukemias. In some cases, the names of the diseases have changed to reflect the additional information. Consequently, cancer registries and pathology departments using the *International Classification of Diseases for Oncology, second edition (ICD-O-2)* have been unable to satisfactorily code these new entities.

Responding to requests for assignment of new code numbers for these entities, in 1998, the International Agency for Research on Cancer (IARC), the cancer division of WHO, gathered a task force to assess whether a revision or new edition of ICD-O was necessary. It was initially thought that only the lymphomas and leukemias would be revised. However, when questionnaires sent to every national registry in the world indicated that new diagnostic terms had been identified in all categories of neoplasms, it was decided to update the entire book. It was not necessary to revise the topography section of ICD-O-2, since it is based on the *International Classification of Diseases, tenth revision.*

In addition, it was desired that the next edition of ICD-O be compatible with other WHO publications such as the series *Histological Typing of Tumors*, known as the Blue Books. In a coordinated effort with the editors of the Blue Books, all terms in existing fascicles were reviewed to ensure that the histologic terminology was included in ICD-O-3. Further, fascicles in preparation have been reviewed to assign ICD-O-3 codes to the terms listed in their type lists.

ICD-O-3, released in December 2000 featured significant changes in the Morphology section. The ICD-O-3 is a dual classification and coding system for both topography and morphology of a neoplasm. The topography code uses the same three- and four-character categories as ICD-10 for malignant neoplasms (C00-C80), allowing greater specificity for the site of nonmalignant neoplasms than is possible in ICD-10. The morphology code describes the specific histologic cell type and its behavior. It indicates the specific histologic term.

***Usage***

The ICD-O-3 is an international dual classification and coding scheme standard for both topography and morphology of neoplasms in tumors and cancer registration and used in pathology laboratories throughout the world.

***Applicability to NCI***

The ICD-O-3 is a neoplasm-specific standard vocabulary to enable collection of information on neoplasms in tumors. The ICD-O-3 is used at NCI and supports Surveillance, Epidemiology and End Result (SEER) reporting and collaboration with international cancer researchers.

***Curation***

ICD-O-3 is included in the NCI Metathesaurus as a local source. In addition, ICD-0-3 terms are incorporated into the NCI Thesaurus, which maintains a more detailed classification of neoplasms required for NCI clinical trial and research needs.

***NCI Role***

The NCI has been a participant in the development of ICD-0-3.

### 2.2.3 International Statistical Classification of Diseases and Related Health Problems (ICD-10, the 10th Revision)

<http://www.cdc.gov/nchs/about/otheract/icd9/icd10cm.htm>

***Sponsor***

The WHO is the United Nations specialized agency for health. WHO's objective, as set out in its constitution, is the attainment by all peoples of the highest possible level of health. The WHO published the International Classification of Diseases and Related Health Problems (ICD-10).

***Description***

The ICD-10 is designed to promote international comparability in the collection, processing, classification, and presentation of mortality statistics. This design includes providing a format for reporting causes of death on the death certificate. The reported conditions are then translated into medical codes using the classification structure and the selection and modification rules contained in the applicable ICD revision. These coding rules improve the usefulness of mortality statistics by giving preference to certain categories, by consolidating conditions, and by systematically selecting a single cause of death from a reported sequence of conditions. The single selected cause for tabulation is called the underlying cause of death, and the other reported causes are the non-underlying causes of death. The combination of underlying and non-underlying causes is the multiple causes of death.

| Revision | Years Covered |
|---|---|
| 1st | 1900-09 |
| 2d | 1910-20 |
| 3d | 1921-29 |
| 4th | 1930-38 |
| 5th | 1939-48 |
| 6th | 1949-57 |
| 7th | 1958-67 |
| 8th | 1968-78 |
| 9th | 1979-98 |
| 10th | 1999-present |

**Exhibit A-1**. List of Revisions to the ICD-10

The ICD-10 is the latest in a series that was formalized in 1893 as the Bertillon Classification or International List of Causes of Death. There is a complete review of the historical background to the classification in Volume 2. While the title has been amended to make clearer the content and purpose and to reflect the progressive extension of the classification scope beyond diseases and injuries, the familiar abbreviation "ICD" has been retained. In the updated classification, conditions have been grouped in a way that was felt to be most suitable for general epidemiological purposes and for the evaluation of health care.

ICD-10 is much larger than ICD-9. Numeric codes (001-999) were used in ICD-9, whereas an alphanumeric coding scheme, based on codes with a single letter followed by two numbers at the 3-character level (A00-Z99), have been adopted in ICD-10. This scheme has significantly enlarged the number of categories available for the classification. Further detail is then provided by means of decimal numeric subdivisions at the 4-character level.

ICD-10, in its entirety, is designed to be a central ("core") classification for a family of disease- and health-related classifications. Some members of the family of classifications are derived by using a fifth or even sixth character to specify more detail. In others, the categories are condensed to give broad groups suitability for use, for instance, in primary health care or general medical practice. There is a multi-axial presentation of Chapter V(F) of ICD-10 and a version for child psychiatric practice and research. The "family" also includes classifications that cover information not contained in the ICD, but having important medical or health implications, e.g., the classification of impairments, disabilities and handicaps, the classification of procedures in medicine, and the classification of reasons for encounters between patients and health workers.

### *Usage*

The ICD was designed as a standard to promote international comparability in the collection, processing, classification, and presentation of mortality statistics. It has been used globally since 1900 on cause-of-death, diseases and nature of injury, and external causes of injury.

### *Applicability to NCI*

Although like the CPT-4, ICD-10 is mostly for billing purposes, it is also a standard means of classifying mortality statistics. For clinical trial data collection and analysis, as well as data sharing and comparing perspectives, ICD-10 should be considered a standard useful to NCI in capture of mortality statistics.

### *Curation*

ICD-10 is included in NCI Metathesaurus.

### *NCI Role*

At their 1996 (Tokyo) meeting, heads of WHO Collaborating Centres for Classification of Disease confirmed the resolution of their 1995 (Canberra) meeting that all proposals for changes to ICD-10 must be sponsored by a Collaborating Centre. Proposals submitted directly to WHO will be returned to the originator with a request that they first be examined by the most appropriate center.

Ten WHO Collaborating Centres for the Classification of Diseases have been established to assist users with problems encountered in the development and use of health-related classifications and, in particular, in using the ICD. The centers meet annually to discuss matters of mutual interest and to advise WHO on the development, implementation, use and revision of health-related classifications. There are three centres for English language users: This process provides for some involvement of NCI in the further development and improvement of the ICD-10.

***Point of Contact***

Ms. Marjorie S. Greenberg
Head, WHO Collaborating Center for the Classification of Diseases for North America
Data Standards, Program Development and Extramural Programs
National Center for Health Statistics
6525 Belcrest Road, Room 1100
Hyattsville, MD 20782
Tel: 1 301 436 4253
Fax: 1 301 436 4233
email: msg1@cdc.gov

Dr. Peter Goldblatt
Office for National Statistics
1 Drummond Gate
London SW1V 2QQ, England
Tel: +44 207 533 5265
Fax: +44 207 533 5103
E-mail: who@ons.gov.uk

Dr. R. Madden
Head, WHO Collaborating Centre for the Classification of Diseases
Director, Australian Institute of Health and Welfare
GPO Box 570
Canberra ACT 2601, Australia
Tel: 61 6 244 1000 (switchboard) 61 6 244 1103 (direct)
Fax: 61 6 244 1111
email: richard.madden@aihw.gov.aus

### 2.2.4 Logical Observation Identifiers Names and Codes (LOINC) <http://www.loinc.org/>

***Sponsor***

The Regenstrief Institute for Health Care has developed LOINC under the sponsorship of various government and private organizations. The Regenstrief Institute for Health Care is a joint enterprise of the Regenstrief Foundation, Inc., the Indiana University School of Medicine, and the Health and Hospital Corporation of Marion County, Indiana. The Institute conducts research to improve health care by optimizing the capture, analysis, content, and delivery of the information needed by patients, their providers and policy makers and conducts interventional studies designed to measure the effect of the application of this research on the efficiency and quality of health care.

***Description***

The LOINC database has two sections: a laboratory portion and a clinical portion. The laboratory portion of the LOINC database contains a universal master file of standard test names and codes for identifying laboratory test results including the categories of chemistry, hematology, serology, microbiology (including parasitology and virology), and toxicology; as well as categories for drugs and the cell counts reported on a complete blood count or a cerebrospinal fluid cell count. Antibiotic susceptibilities are a separate category. The clinical portion of the LOINC database includes entries for vital signs, hemodynamics, intake/output, electrocardiogram (EKG), obstetric ultrasound, cardiac echo, urologic imaging, gastroendoscopic procedures, pulmonary ventilator management, selected survey instruments, and other clinical observations. The purpose of the LOINC laboratory and clinical test results code set is to facilitate the exchange and pooling of results for clinical care, outcomes management, and research. Each LOINC record corresponds to a single test result or panel. The record includes fields specifying:

1. Component (analyte)
2. Property measured
3. Timing
4. Type of sample
5. Type of scale
6. Method (optional)

Additional information may include lab test short names, related words, synonyms, and comments for all observations.

The following data sources were used to develop the LOINC dataset: Silver Book for the International Union of Pure and Applied Chemistry (IUPAC) and the International Federation of Clinical Chemistry (IFCC), textbooks of clinical pathology, expertise and work of LOINC members, EUCLIDES, and master files from Indiana University/Regenstrief, University of Utah, Association of Regional and University of Pathologists (ARUP), Mayo Medical Laboratories, LDS Hospital in Salt Lake City, the Department of Veterans Affairs, Quest Diagnostics, and the University of Washington. The database includes over 30,000 observation concepts. The list was first released in April 1996. There have been thirteen revisions with the last release in October 2003.

LOINC databases are available in a number of file formats: ACCESS, tab delimited ASCII, and PDF. The Regenstrief LOINC Mapping Assistant (RELMA) is a Windows-based mapping utility for utilization of LOINC information. This information is publicly available on the LOINC Web site, <http://www.loinc.org/>.

LOINC codes have been adopted and/or endorsed by a number of organizations including the College of American Pathologists (CAP), the American Clinical Laboratory Association (ACLA), Quest Diagnostics, Lab Corp, SmithKline Beecham, the Associated Regional and University Pathologist (ARUP), Mayo Medical Laboratories, University of Colorado, Intermountain Health Care, Kaiser Permanente, Clarian Health, Partners Healthcare System of Boston, Care Group of Boston, Mayo Medical Group, and the Department of Defense (DoD). All U.S. veterinary medical laboratories have committed to using LOINC. Empire Blue Cross and Aetna Health Care have adopted LOINC for internal purposes. Internationally, LOINC is being used in Switzerland, Canada, and Germany. The National Library of Medicine has included LOINC codes in the Unified Medical Language System (UMLS) and they have been proposed for inclusion in the Health Insurance Portability and Accountability Act (HIPAA) electronic attachment specifications. The CDC is using the codes for electronically reporting/transmitting communicable disease information. The HHS, DoD, and the Veterans' Health Administration (VHA) have also adopted this coding system for exchange of clinical laboratory results.

### *Applicability to NCI*

LOINC codes are endorsed by the CHI Initiative, and could be adopted for description of laboratory tests. The new version of LOINC has combined laboratory test names and health care related data elements/form questions as the components. Intelligent manipulation of the database will be required to extract only the lab test results domains for NCI usage. Over time, LOINC codes could be used to identify a wide range of the clinical observations collected in the course of NCI sponsored clinical trials. In the past, the LOINC committee has been open to the addition of new codes to meet emerging requirements.

### *Curation*

LOINC information is being registered in the caDSR and is available in the EVS on both the NCI Distributed Terminology Server and the NCI Metathesaurus server. LOINC may be registered in caDSR as individual components for Component (analyte), Property Measured, Timing, Type of Sample, Type of Scale, and Methods; as the fully specified laboratory test names/codes; and/or a link to the RELMA application could be provided so that users could query for desired values.

### *NCI Role*

NCI staff should participate in the development of this data standard to ensure that the standard meet the needs of the agency for describing clinical laboratory tests. This will involve monitoring the LOINC listserv, reviewing updates of the standard, and participation in periodic meetings. Frequent updates to the standard are issued (13 revisions since the first release in April 1997) that will need to be registered in a timely fashion.

### *Point of Contact*

LOINC
c/o Regenstrief Institute
1050 Wishard Boulevard
Indianapolis, IN 46202
loinc@regenstrief.org

### 2.2.5 The Medical Dictionary for Regulatory Activities (MedDRA) <http://www.ich.org/ichMedDRA.html/>

***Sponsor***

MedDRA has been developed under the sponsorship of the International Committee on Harmonization (ICH). The International Federation of Pharmaceutical Manufacturers Associations (IFPMA) holds the copyright for MedDRA. A Maintenance and Support Services Organization (MSSO) has been created to serve as the repository, maintainer, and distributor of MedDRA. The MSSO is expected to ensure the maintenance of the integrity and medical correctness of the terminology. Responsibility for establishing and operating the MSSO was awarded to Northrop Grumman - a leading U.S. systems integrator and information technology consulting company. The MSSO has implemented an ongoing terminology maintenance process, accepting change requests from subscribers, and proactively proposing changes to MedDRA. The MSSO is also responsible for terminology distribution, the MedDRA User Group, and providing help desk support.[6]

***Description***

MedDRA has been described as "pragmatic, clinically validated medical terminology with an emphasis on ease-of-use data entry, retrieval, analysis, and display, with a suitable balance between sensitivity and specificity, within the regulatory environment."[7] MedDRA was developed as a clinically validated terminology for use throughout the regulatory process. "The developers of the terminology designed a structure that promotes specific and comprehensive data entry and flexible data retrieval."[8] The following categories of medical data are included:

- Signs
- Symptoms
- Diseases
- Diagnoses
- Therapeutic indications
- Names and qualitative results of investigations, including pharmacokinetics
- Surgical and medical procedures
- Medical/social/family history

From a structural perspective, the terminology supports three kinds of relationships between terms:

- Equivalence: Synonymous terms are grouped under the category of "Preferred Terms."
- Hierarchical Association: Terms are grouped into the following list of categories (listed from general to specific) System Organ Class (SOC), High Level Group Term (HLGT), High Level Term (HLT), Preferred Term (PT), Lowest Level Term (LLT). The reader should note that each LLT is linked to a single PT - it may be a synonym, lexical variant, or sub-element of the parent PT. Each PT is linked to a primary SOC, however it may also be linked to additional SOCs. The HLGT, and HLT serve as groupers for PTs. However, each grouping is formed in the context of a single SOC.

[6] The reader should note that the active role of a commercial MSSO strongly affects the fee structure for using MedDRA.

[7] < http://www.meddramsso.com/ >, FAQ

[8] MedDRA Introductory Guide v6.1 Page 5, MSSO-DI-6003-6.1.0, September 2003.

- Associative Grouping: "Associative groupings of terms are linked horizontally in the terminology. Special Search Categories (SSCs) link terms that are neither equivalent nor hierarchically related. Instead, the terms may be related symptoms, signs, or diseases relevant to a diagnosis (e.g., IMMEDIATE HYPERSENSITIVITY AND ANAPHYLACTIC REACTIONS (SSC), ARREST (CARDIAC) (SSC), or BONE MARROW DEPRESSION (SSC)."[9]

***Usage***

As previously noted, MedDRA was developed to support the requirements of pharmaceutical regulatory and adverse event reporting. It has the solid support of the international pharmaceutical industry as the terminology to be used in regulatory reporting. In the United States, the FDA has supported reporting using MedDRA since 1997 and is required to use MedDRA for regulatory reporting by international agreement. A notice of proposed rulemaking to mandate adverse drug event reporting using MedDRA has been published by the FDA.

***Applicability to NCI***

MedDRA is the current primary terminology standard for adverse event reporting.

***Curation***

NCI is a MedDRA Core subscriber and MedDRA is available in the NCI Metathesaurus. Also EVS supplies MedDRA as a file set suitable for loading into database schema. MedDRA codes could also be evaluated for registration as a value domain in the caDSR.

***NCI Role***

MedDRA is designed for use as a standard for regulatory reporting in the pharmaceutical arena and is maintained by a commercial organization. Northrop Grumman (the MSSO) solicits MedDRA subscribers for proposed changes to the vocabulary. Each subscriber has the right to request a finite number of changes, but may request additional changes for a fee. These proposals are evaluated by an international panel of medical personnel, and reviewed for quality by the MSSO. As a MedDRA subscriber, NCI is in a position to request changes. Currently EVS distributes announcements of proposed changes to a mailing list of NCI personnel.

### 2.2.6 National Cancer Institute (NCI) Thesaurus
<http://nciterms.nci.nih.gov/>

***Sponsor***

The National Cancer Institute has developed NCI Thesaurus as part of NCI's Enterprise Vocabulary Services (EVS), combining and extending earlier NCI terminologies and coding systems to provide a reference terminology for use by NCI and the wider cancer community. The Institute conducts and supports a wide range of research, clinical, and public information activities, and has designed NCI Thesaurus to provide a comprehensive, science- and logic-based framework for meeting coding and retrieval needs including translational integration of basic and clinical research.

[9] Ibid. Page 7

### *Description*

NCI Thesaurus provides cancer-related controlled terminology within a concept-based, description-logic framework that characterizes the meanings of and relationships between concepts, providing rich synonymy, definitions, internal and external codes and cross-references, and other features including the history of any changes made to each concept. The current set of over 36,000 concepts includes some 100,000 synonyms, acronyms, codes, and other identifiers, and is rapidly expanding to meet the needs of its users. These concepts are organized in 20 distinct "Is-A" hierarchies and interconnected by 89 types of description logic "role" relationships, with roughly 60,000 defined and inferred concept-to-concept relationships.

Areas of special interest include:

- Neoplasms: Detailed coverage of 6,000 distinct neoplastic categories, characterized by about 25,000 role relationships covering anatomy, tissue, cell, cytogenetic, molecular, clinical, and other features.
- Cancer-related diseases, disorders, findings, and abnormalities: About 5,000 concepts used to characterize cancer or for supportive care or adverse event reporting.
- Anatomy: A CHI-recommended standard with 4,400 concepts including unique coverage of microanatomy.
- Genes and gene products: Over 2000 genes and 2,300 corresponding proteins and other gene products chosen for their known involvement in neoplastic disease, or for their similar structure or function to these genes; characterized by more than 12,000 role relationships to biological processes, biochemical pathways, disease, and anatomy.
- Chemicals, drugs, and combination therapies: Over 6,700 concepts include comprehensive coverage of agents used in both open and closed cancer clinical trials, including health professional definitions for most agents in open trials.
- Experimental models: Over 1,000 concepts developed in collaboration with the Mouse Models of Human Cancers Consortium (MMHCC), together with initial coverage of other models.

NCI Thesaurus was initialized five years ago with Physician Data Query (PDQ) Terminology and other NCI vocabulary, then greatly extended and revamped by an interdisciplinary team drawing on current scientific literature and databases, comparison with and mapping to other biomedical terminologies in the NCI Metathesaurus, project and working groups in diverse areas, and detailed review by outside experts including the College of American Pathologists (CAP).

### *License*

It is freely available internationally under an open content license.

### *Usage*

Currently in use in a variety of NCI programs and projects, including caDSR/caCORE and NCI's PDQ and Web public information services. It is distributed through the caCORE, as UMLS Rich Release Format text files, and in Ontology Web Language (OWL) format; the OWL version is gaining use within the Protégé and Semantic Web communities.

### *Applicability to NCI*

Direct, as a resource developed primarily to meet NCI needs.

***Curation***

EVS is a partnership between the NCI Office of Communications and the NCI Center for Bioinformatics. The EVS Project facilitates the standardization of vocabulary across the Institute and the larger cancer biomedical domain. In addition to formal governance, provided by the NCI Vocabulary Executive Group and outside experts, collaboration with current and prospective users of EVS services is a major emphasis. At present a number of NCI and affiliated organizations are ongoing collaborators in defining EVS content. In addition, NCI EVS is active in standard development organizations, interagency efforts to develop public domain vocabulary products, and software. As opportunities arise, the EVS Project engages in NCI-wide harmonization initiatives addressing how vocabulary is used to code, retrieve and aggregate information.

### 2.2.7 North American Association of Central Cancer Registries, Inc. (NAACCR, Inc.)

<http://www.naaccr.org/index.asp>

***Sponsor***

Established in 1987, NAACCR, Inc. is a collaborative umbrella organization for cancer registries, governmental agencies, professional associations, and private groups in North America interested in enhancing the quality and use of cancer registry data. All central cancer registries in the United States and Canada are members. The National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) Program, is a sponsoring organization for this project.

***Description***

The NAACCR, Inc. is a professional organization that develops and promotes uniform data standards for cancer registration; provides education and training; certifies population-based registries; aggregates and publishes data from central cancer registries; and promotes the use of cancer surveillance data and systems for cancer control and epidemiologic research, public health programs, and patient care to reduce the burden of cancer in North America. This enables compilation of case-specific data into useful and meaningful information, and facilitates comparison of data across different registries.

When NAACCR, Inc. was first organized, it focused its efforts to achieve consensus on cancer registration standards among the many standard setters in the United States and Canada. They include the American College of Surgeons, the National Cancer Institute, and the Canadian Cancer Registry at Statistics Canada. Today nearly all registries throughout the United States and Canada have adopted the NAACCR consensus standards. These standards are updated annually. Maintaining current standards to meet the needs of the NAACCR community is an ongoing and major NAACCR activity. NAACCR publishes a record layout for transmission of cancer data. The layout is based on a number of existing standards, which are incorporated by reference, including the standards of the World Health Organization and the NCI's SEER program. The layout has some limitations in data names, as NAACCR has its own EDITS software that is used to analyze data collected.

***Applicability to NCI***

NAACCR incorporates data standards for Demographic, Tumor and Staging, Treatment and Follow-up, Patient Identifiers and Physicians, Sender Identification, Record Identification, Corrections, and Group Recodes/Conversions. In some cases, NAACCR publishes differences between the reporting standards of the participating standards organizations. Of particular interest to NCI may be the American Joint Committee on Cancer (AJCC). The *AJCC Cancer Staging Manual* provides a guide to the TNM (Tumor, Nodes, Metastasis) staging system, a classification scheme for cancer of 47 anatomic sites that describes the primary tumor, regional lymph nodes, and metastasis. The staging scheme is useful to physicians and health care practitioners to determine the extent of disease. The stage is then used to guide the management of cancer patient care.

The standards incorporated by NAACCR should be reviewed in detail to determine applicability to NCI efforts. Detailed coding instructions for many data items in the data exchange record are implied by the "Source of Standard" located in NAACCR Standards for Cancer Registries Volume II *Data Standards and Data Dictionary*. The following list includes the current reference manuals:

- *AJCC Cancer Staging Manual Sixth Edition (TNM).*
- *Canadian Cancer Registry Data Dictionary.*
- *COC Facility Oncology Registry Data Standards (FORDS).*
- *Collaborative Staging Manual and Coding Instructions, Version 1.0.*
- *NAACCR Standards for Cancer Registries Volume I: Data Exchange Standards and Record Description.*
- *NAACCR Standards for Cancer Registries Volume II: Data Standards and Data Dictionary.*
- *SEER Program Code Manual Version 3.1.*
- *SEER Extent of Disease - 1988: Codes and Coding Instructions.*
- *SEER Summary Stage 2000.*
- *WHO ICD-O Third Edition.*

### *Curation*

NAACCR needs to be viewed as a compilation of standards into a data exchange format. Its component parts are being reviewed to determine potential standard data elements of interest. Much of the controlled terminology used by SEER, including AJCC stage terms and ICD-03, are already incorporated into NCI Thesaurus. ICD-03 is also included in the NCI Metathesaurus as a local vocabulary source.

### *NCI Role*

NCI EVS is currently collaborating with SEER on enhancement and compatibility of the SEER data dictionary with NCI Thesaurus.

### *Point of Contact*

NCI Point of Contact: Benjamin Hankey, Chief Cancer Statistics Branch, bh43a.@nih.gov, 301-402-5288

North American Association of Central Cancer Registries, Inc.
2121 West White Oaks Drive, Suite C

Springfield, Illinois 62704
Phone: (217) 698-0800
Fax: (217) 698-0188

### 2.2.8 Systematized Nomenclature of Human and Veterinary Medicine (SNOMED)
<http://www.snomed.org/>

***Sponsor***

SNOMED International, a division of the College of American Pathologists (CAP), is the developer of SNOMED. CAP is a medical society serving nearly 16,000 physician members and the laboratory community throughout the world. It is the world's largest association composed exclusively of pathologists and is widely considered the leader in laboratory quality assurance.

***Description***

SNOMED has been described as "A comprehensive set of concepts, terms and codes used by physicians, dentists, nurses, allied health professionals, veterinarians, and others." It is a reference terminology - "a set of concepts and relationships that provide a common reference point for comparison and aggregation of data about the entire health care process, recorded by multiple different individuals, systems, or institutions."[10] - that is intended to be used for coded representations of patient diagnostic and clinical information. Such a reference terminology is considered to offer the following benefits:

- A representation of clinical meaning to which various user interfaces and systems can refer for semantic interoperability.
- The "semantic scaffolding" onto which the various terms and concepts of health care can be placed, making the concept meanings and interrelationships explicit
- A structure that facilitates reliable machine generated interpretation and aggregation of data.[11]

SNOMED is a multi-axial terminology system; that is to say it is composed of multiple semantic hierarchies, each addressing a separate aspect of a medical concept.

A SNOMED concept can be drawn from a single axis, or it may represent a combination of atomic concepts from multiple axes. For example: "A diagnosis in SNOMED may consist of a topographic code, a morphology code, a living organism code, and a function code. When a well-defined diagnosis for a combination of these four codes exists, a dedicated diagnostic code is defined. For example, the disease code D-13510 (Pneumococcal pneumonia) is equivalent to the combination of:

- T-28000 (topology code for Lung, not otherwise specified),
- M-40000 (morphology code for Inflammation, not otherwise specified), and
- L-25116 (for Streptococcus pneumoniae) along the living organism axis.

[10] Spackman KA, Campbell KE, Cote RA. SNOMED RT: A reference terminology for health care. In Masys DR. (Ed): *Proceedings of the 1997 American Medical Informatics Association Fall Symposium* 1997. Philadelphia, Hanley & Belfus: 640-644.

[11] Michael Stearns, Representing Clinical Data in the Healthcare Enterprise, PowerPoint presentation, 2000.

Tuberculosis (D-14800), for instance, could also be coded as Lung (T-28000) + Granuloma (M-44000) + Mycobacterium tuberculosis (L-21801) + Fever (F-03003). However, this can be confusing since tuberculosis is not only restricted to the lung."[12]

SNOMED was first published in 1975 and has been updated and extended since that time. This process of extension has also been accompanied by changes of name. Version 3.5 issued in 1998 was released as SNOMED International. In 2000, a new version was released as SNOMED RT. This version was intended to "allow for the full integration of all medical information in the electronic medical record into a single data structure, facilitating interoperability between a wide variety of systems and clinical records."[13] The most recent release, SNOMED CT is based on collaboration between CAP and the UK National Health Service that combines SNOMED RT with the Clinical Terms Version 3 of the NHS thesaurus of health care terms (Read Codes). This updated version was described - as of January 2003 - as containing 344,000 concepts with unique meanings and formal logic-based definitions organized into 19 hierarchies. The fully populated table contains more than 913,000 English language descriptions or synonyms.

### *Usage*

SNOMED has been developed to be widely used to codify medical records and other computer-based patient records. However, actual use of SNOMED has been limited to a small number of hospitals and to a few multi-hospital systems such as Kaiser Permanente. There are two primary reasons for this lack of acceptance and use. In the first place, using a coding system such as SNOMED entails comprehensive and painful procedural changes for the health care provider. These changes include the inculcation and/or enforcement of consistent documentation and coding practices among medical professionals, and are linked to the implementation and use of electronic health record systems - whose purchase and installation is also a major task. Secondly, the licensing fees for SNOMED have been perceived as both substantial and excessive. This has led not only to an unwillingness to license SNOMED, but to a widespread reluctance to use SNOMED as a key architectural element with health care information systems due to fears that future costs will increase.

In the United States, this second barrier to using SNOMED appears to have been removed by recent U.S. government action. CAP has signed a 5-year sole source contract with the National Library of Medicine to make SNOMED CT core content available free-of-charge to "U.S. federal agencies, state and local government agencies, territories, the District of Columbia, and any public, for-profit and non-profit organization located, incorporated and operating in the U.S."[14] It is expected that “The investment of federal resources at the national level will make the use of SNOMED CT in electronic patient records more affordable and easier to implement for many more American health care organizations. When built into electronic patient record products, SNOMED CT enables primary and specialty care providers and patients to share comparable data at any time, from any place. This ability has the potential of greatly reducing medical errors associated with traditional paper records."[15] It seems likely that this action will indeed

[12]

Handbook of Medical Informatics, Website, v3.3,
Editors: *J.H. van Bemmel*, Erasmus University, Rotterdam, *M.A. Musen*, Stanford University, Stanford. Glossary.

[13] < http://www.snomed.org/snomedrt_txt.html>

[14] SNOMED International, Press Release, July 1, 2003.

[15] Ibid.

substantially increase the potential for health care providers to evaluate migration to using SNOMED for coding medical concepts.

***Applicability to NCI***

Use of SNOMED potentially offers substantial benefits to the NCI. While not currently in wide use at health care sites, SNOMED's broad coverage of the health care domain increases its potential as a widely used standard. CHI has recommended that SNOMED be used as the standard for anatomy and pathology (as is NCI Thesaurus), diagnosis and problem lists, interventions and procedures, Lab Result contents, and nursing. NCI will continue to evaluate SNOMED and other vocabularies to meet its needs. For example, NCI already has applications that require finer granularity for neoplasm diagnosis than SNOMED provides.

It is important to keep in mind that SNOMED is an extremely broad system that is explicitly designed to include all relevant medical concepts. However, its use will take place in a specific context. For example, in some situations the diagnosis for a patient is relevant, in others it will be the medication that was administered. The use of SNOMED should be tailored to the particular situation. That is to say, the documented procedures and information systems involved should make clear the context within which vocabulary is being used, and limit the available and valid concepts accordingly.

***Curation***

SNOMED CT is included in both the NCI Metathesaurus, where it is mapped to other controlled terminologies including NCI Thesaurus, and as a stand-alone source in the EVS DTS browser. Utilization will be most productive if it is possible to provide links between particular SNOMED axes, or constellations of axes, and the context in which they are to be used.

***NCI Role***

SNOMED is developed and controlled by the College of American Pathologists through its SNOMED International division. In addition to being a licensee and having representation on the SNOMED editorial board, NCI has worked cooperatively with the CAP and SNOMED staff on reviews of NCI Thesaurus neoplasm terminology. SNOMED and NCI participate jointly in a number of standards organizations, including HL7.

## 2.3 Genomics

### 2.3.1 Gene Ontology (GO) Consortium

< http://www.geneontology.org/>

***Sponsor***

> "The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products in different databases. The project began as a collaboration between three model organism databases: FlyBase (Drosophila),the Saccharomyces Genome Database (SGD) and the Mouse Genome Database (MGD) in 1998. Since then, the GO Consortium has grown to include many databases, including several of the world's major repositories for plant, animal and microbial genomes."

***Description***

> "The GO collaborators are developing three structured, controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. There are three separate aspects to this effort: first, we write and maintain the ontologies themselves; second, we make associations between the ontologies and the genes and gene products in the collaborating databases, and third, we develop tools that facilitate the creation, maintenance and use of ontologies."

Terms in molecular functions describe activities at the molecular level, such as adenylate cyclase activity or chromatin binding, while a biological process, such as protein biosynthesis, is accomplished through "ordered assemblies of molecular functions". Terms in cellular components describe the sub-cellular locations where gene products reside. Annotations (associations) of genes and gene products to GO terms by expert curation teams using various computational tools and expertise, in essence provide a concise summary of relevant literature about gene product function.

The GO terms are connected in a network that can be represented by a Directed Acyclic Graph (DAG), meaning that each term (or node) can have multiple parent and child relationships (or edges), but no path can start and end at the same node. Such a structure allows for the simplification of many graph calculations, and maintains a reasonable representation of the complex biological domain knowledge. The rapid development and adoption of GO, and availability of associations of GO terms to genes and their products have made GO the de facto standards for genome annotation.

As GO expanded rapidly to include controlled vocabulary terms that cover as much biological concepts and knowledge content, relatively less attention was paid to the development of ontological elements and rules as defined by information science and philosophy, which emphasize logic frameworks between objects, processes, and relationships. The trade-offs of the GO development process are less ontological coherence, stability and scalability, and difficulties in developing software tools for validation, maintenance, and automated reasoning.

Although similar functional annotation efforts, such as Swissprot keywords, EC, TIGR roles, and MultiFun, have partially overlapping content and terminology with it, GO has largely been adopted as the standard by most genomic databases.

### *License*

The GO database, vocabularies, and annotations provided by member organizations are in the public domain.

> "The GO Consortium gives permission for any of its products to be used without license for any purpose under three conditions: 1. That the Gene Ontology Consortium is clearly acknowledged as the source of the product; 2. That any GO Consortium file(s) displayed publically include the version number(s) and/or date(s) of the relevant GO file(s); 3. That neither the content of a GO file(s) nor the logical relationships embedded within the GO file(s) be altered in any way."

### *Usage*

The GO vocabularies are widely adopted for gene annotations by the major genome databases, such as the National Center for Biotechnology Information ( NCBI), Swissprot (UniProt), Ensembl, European Bioinformatics Institute (EBI), The Institute for Genomics Research (TIGR), etc. Such annotations are integrated in various browsers of genomic information, e.g. NCBI Locuslink (Entrez Gene) query display, and the Ensembl genome browser. These browsers allow the user to lookup function(s) performed by a particular gene product as annotated to relevant GO terms, and other gene products in the same or different species that have similar functional annotations.

The GO consortium also created the GO term and gene association database. GO, as well as other organizations, have made software and software development tools available for bioinformatics developers, such as QuickGO (EBI), GoMiner (NCI), and CGAP GO (NCI) Browser, etc.

***Applicability to NCI***

The GO terms are effectively the de facto standards for genome annotation. GO is mapped in the NCI Metathesaurus and is hosted as a standalone vocabulary in EVS, made accessible through the API and browser. Gene entities in the NCI Thesaurus have been annotated with GO terms using a mapping table. Although GO contains many shortcomings in its design, there are few alternatives for gene product annotation that rival the scope, coverage, and utilities of GO. Therefore it is recommended that NCI should continue to make GO available for application development.

***NCI role***

NCI will continue to update the GO in EVS, and monitor its value as a gene vocabulary for gene annotations.

### 2.3.2 Human Genome Organization (HUGO) Gene Nomenclature Committee (HGNC)

<http://www.gene.ucl.ac.uk/nomenclature/>

***Sponsor***

The HGNC is a group based in the Department of Biology, University College London, UK with funding from the National Institute of Health (US) and the Medical Research Council (UK).

***Description***

The HGNC approve a gene name and symbol for each known human gene, to make sure it is unique. There are now 19408 approved gene symbols, 20531 aliases, and 4090 withdrawn symbols as of 05/28/2004.

> "Individual new symbols are requested by scientists, journals (e.g. Genomics, Nature Genetics and Cytogenetics and Cell Genetics) and databases (e.g. RefSeq, OMIM, GDB, MGD and LocusLink), and groups of new symbols by those working on gene families, chromosome segments or whole chromosomes. As the human genome sequence analysis nears completion there is an increasing demand for the rapid approval of gene symbols. However, in all cases considerable efforts are made to use a symbol acceptable to workers in the field."

The HGNC also mapped the gene symbols to identifiers used by a number of public databases, such as RefSeq, LocusLink, GDB, SWISSPROT, OMIM. The same data can be obtained from LocusLink.

***License***

As a funding condition by NIH and the Medical Research Council (MRC), the HGNC nomenclature is freely available to all.

***Usage***

The HGNC nomenclature provides gene names and symbols for human genes, which should be used as controlled terminology in data representations of such.  It is currently used by major human genome databases such as NCBI, Ensemble, UCSC genome browser, etc.

***Applicability to NCI***

Human Gene Nomenclature (HUGN) sponsored by the Human Genome Organization (HUGO) is a recommended standard by the Consolidated Health Informatics Initiative (CHI).

***NCI role***

NCI staff has played a role in evaluating standards for the federal health care sector to exchange information regarding the role of genes in biomedical research and healthcare, using a single unambiguous genetic nomenclature for the CHI.

### 2.3.3 The Microarray Gene Expression Data (MGED) Society

<http://www.mged.org/>

***Sponsor***

The MGED society was founded in 1999 by major microarray users and developers including Affymetrix, Stanford University and The European Bioinformatics Institute (EBI).  The goal of the MGED Society is "to facilitate the sharing of microarray data generated by functional genomics and proteomics experiments".  "The current focus is on establishing standards for microarray data annotation and exchange, facilitating the creation of microarray databases and related software implementing these standards, and promoting the sharing of high quality, well annotated data within the life sciences community."

***Description***

The MGED Society has developed standards and is developing extended standards on exchange of microarray data.  The defined standards are "MIAME - Minimum Information About a Microarray Experiment - a document which outlines the minimum information that should be reported about a microarray experiment to enable its unambiguous interpretation and reproduction" (version 1.1 as of 05/2004), and MicroArray and Gene Expression (MAGE). "MAGE consists of three parts: An object model (MAGE-OM), a document exchange format, which is derived directly from the object model (MAGE-ML), and software toolkits (MAGE-stk), which seek to enable users to create MAGE-ML" (version 1.1 as of 05/2004).  In addition, the Ontology Working Group of the MGED Society has also developed a MGED Ontology (MO, version 1.1.8 as of 05/2004). The emerging standards include MIAME extensions for toxicogenomics, and those for exchanging information on data transformation and normalization.

### *Minimum Information About a Microarray Experiment (MIAME)*

MIAME defines the minimum information that one should include when exchanging microarray data. It contains two main sections: one on the array design, the other on the experiments. The array design section specifies information on the array platform (spotted or in situ), surface and coating specification, type of reporter (oligos or PCR products), etc; while the experiment section specifies information on experimental design (normal vs disease, control vs treatment, time course), sample preparation, hybridization parameters (labels, blocking agents, number of washes), data collection and processing (instrument parameters, image analysis software, normalization algorithms and error models).

### *MicroArray and Gene Expression – MAGE*

MAGE is an umbrella covering three related parts: the MAGE data exchange object model (MAGE-OM), the data exchange format in markup language (MAGE-ML) implemented in XML, and the software development toolkit, MAGEstk.

MAGE-OM is a platform-independent specification that defines a data model, as a set of packages containing objects or classes, for microarray gene expression data exchange. It has been submitted and maintained by the Object Management Group (OMG), a non-profit consortium that produces and maintains computer industry specifications. For example, according to MAGE-OM, information about a particular hybridization should be stored as a hybridization object (class) within the BioAssay package; while the numbers of probes on an array and on a row of a particular array are stored as attributes to the ArrayDesign and ZoneLayout classes respectively within the ArrayDesign package.

“Microarray Gene Expression Markup Language (MAGE-ML) is a language designed to describe and communicate information about microarray based experiments. MAGE-ML is based on XML and can describe microarray designs, microarray manufacturing information, microarray experiment setup and execution information, gene expression data and data analysis results.” For example, a particular probe on an Affymetrix array can be coded in MAGE-ML as <Reporter_ref identifier=”U39447_at”></Reporter_ref> and the data associated with the probe on a particular chip can be coded as <Datum value=”0.485110”></Datum>. The ability to read and write MAGE-ML has been and is being incorporated into major microarray databases, and analysis tools. These efforts are aided by the software development tool kit, MAGEstk, offered by the MGED Society.

Some of the difficulties of using MAGE-ML include its complexity, the resultant large file size when all elements are XML tagged, and vendor specific implementations so that the same object may be coded with slightly different XML syntax by different vendors.

### *MGED Ontologies*

The MGED Society has supplemented the MIAME and MAGE standards with an ontology to provide standard terms or vocabularies so that data from different sources can be queried using the same syntax. “The primary purpose of the MGED Ontology is to provide standard terms for the annotation of microarray experiments. These terms will enable structured queries of elements of the experiments. Furthermore, the terms will also enable unambiguous descriptions of how the experiment was performed." The terms are organized into classes and provided in the form of an ontology, which is "a formal and declarative representation which includes the vocabulary (or names) for referring to the terms in that subject area and the logical statements that describe what the terms are, how they are related to each other, and how they can or cannot be related to each other." For example, a fragment of DNA generated by a PCR reaction may be called "PCR product", "PCR fragment", or "PCR amplicon" by different laboratories, and the MGED ontology intends to unify the expression to "PCR amplicon". The terms included in the MGED ontology are generated by domain experts both within and outside the MGED Society.

***License***

All MGED standards are "open source" and downloadable from links provided from the MGED website. In addition, the MGED Society provides documentation on using the standards and encourages development of open source software tools using the standards.

***Usage***

The MIAME, MAGE, and MGED ontology are still maturing and have been adopted slowly by the microarray databases and tool makers, such as ArrayExpress (EBI), Gene Expression Data Portal (NCI), Stanford Microarray Database, SAS Microarray Solution, Rosetta Resolver, etc. The MIAME standards are being used as a guide for the kind of information that should be included when storing and exchanging microarray data. MAGE-OM is used to map relational database implementations of microarray data to objects for MAGE-ML data exchange, and as a guide for new database design. The availability of the MGED ontology should help facilitate the adoption of the MGED standards, and realize the benefits envisioned by the MGED Society.

***Applicability to NCI***

The NCI cancer array informatics project, caArray for microarray experiments is one of the first microarray databases to embrace the MGED standards, in being MIAME compliant, adhering to MAGE-OM, and providing MAGE-ML formatted data. The NCI is using the MGED Object Model, exchange format, and ontologies. Therefore, this standard is currently applicable to NCI.

***Curation***

The MGED ontology is already included in EVS.

***NCI role***

NCI has adopted the MGED Ontology, and has collaborated with the MGED Ontology Working Group to expand the MGED-O to better describe cancer- and clinical-related areas and utilize the NCI Thesaurus to describe these where appropriate. The NCI maintains an ongoing collaboration with the MGED OWG.

### 2.3.4 Mouse Anatomy (adult – MA, and development – EMAP)

<http://www.informatics.jax.org/searches/anatdict_form.shtml>

***Sponsor***

The mouse anatomy projects are community-supported efforts with collaboration between the primary sponsors: the Jackson Laboratory, the University of Edinburgh, and the MRC Human

Genetics Unit. “Stages 1 through 26 (embryonic development) of the Standard Anatomical Nomenclature Database are being developed at the Department of Biomedical Sciences, University of Edinburgh, Scotland and the MRC Human Genetics Unit, Edinburgh, as part of The Mouse Atlas and 3D Graphical Gene Expression Database Project. Stages 27 and 28 (newborn and postnatal mouse) are being developed by the Gene Expression Database Project at The Jackson Laboratory.”

***Description***

The mouse anatomy ontologies contain 28 controlled vocabularies corresponding to the stages of embryonic and postnatal development of the mouse. Within each vocabulary are terms that enumerate the various anatomical structures arranged in a hierarchical structure. For example, “atrium” is part of “heart”, which is part of “cardiovascular system”, which is part of “organ system”, which can be part of “Stage 28” or other embryonic stages, such as “Stage 26: embryo”.

Although the mouse anatomy vocabularies and the associated development for human anatomy have a lot of overlap in content with medically derived standards such as MESH, UML, and NCI Metathesaurus, the anatomy ontologies are developed through a similar process as GO, which makes it easy to adapt tools developed for GO to utilize anatomy vocabularies.

***License***

The Mouse Anatomy ontologies are parts of the community supported Open Biological Ontologies (OBO) projects, and are “open and can be used by all without any constraint other than that their origin must be acknowledged and they cannot be altered and redistributed under the same name.”

***Usage***

The mouse anatomy terms are being used by the Gene Expression Database (GXD) at the Jackson Laboratory to annotate gene products that have been determined to be expressed in various anatomical structures at different developmental stages. These terms are also used to annotate anatomical images by the Edinburgh Mouse Atlas Project (EMAP). Collaboration between EMAP and GXD would allow integration of 3D mouse anatomy atlas with gene expression data.

***Applicability to NCI***

Since a mouse-related terminology is widely used in NCI cancer related research, the anatomy ontologies should be useful for data integration. It is likely to be incorporated into NCI’s EVS.

***NCI role***

NCI will participate as a collaborator.

### 2.3.5 Taxonomy
<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=taxonomy>

***Sponsor***

National Center for Biotechnology Information (NCBI) creates a number of public molecular biology databases including the taxonomy database.

***Description***

"The NCBI taxonomy database contains the names of all organisms that are represented in the genetic databases with at least one nucleotide or protein sequence." It is important to note that "The NCBI taxonomy database is not a primary source for taxonomic or phylogenetic information. Furthermore, the database does not follow a single taxonomic treatise but rather attempts to incorporate phylogenetic and taxonomic knowledge from a variety of sources, including the published literature, web-based databases, and the advice of sequence submitters and outside taxonomy experts. Consequently, the NCBI taxonomy database is not a phylogenetic or taxonomic authority and should not be cited as such."

***License***

All databases created by NCBI including the Taxonomy database are public domain information.

***Usage***

The NCBI taxonomy database is used by all the major genome databases such as NCBI, Ensembl, UCSC genome browser, Swissprot, etc as organism definition and terms.

***Applicability to NCI***

Links are provided at the concept level to these sources through NCI Thesaurus.

***NCI role***

None current.

## 2.4 Drug Identification

### 2.4.1 National Drug File Reference Terminology (NDF-RT)

<http://www.aspe.hhs.gov/sp/nhii/Standards.html>

***Sponsor***

The mission of the Veterans Healthcare System (VHS) is to serve the needs of America's veterans by providing primary care, specialized care, and related medical and social support services.

The Veterans' Health Administration (VHA) initiated the National Drug File (NDF) Reference Terminology (RT) project as part of an effort to standardize all medical terminologies within the Department of Veterans Affairs (VA).

***Description***

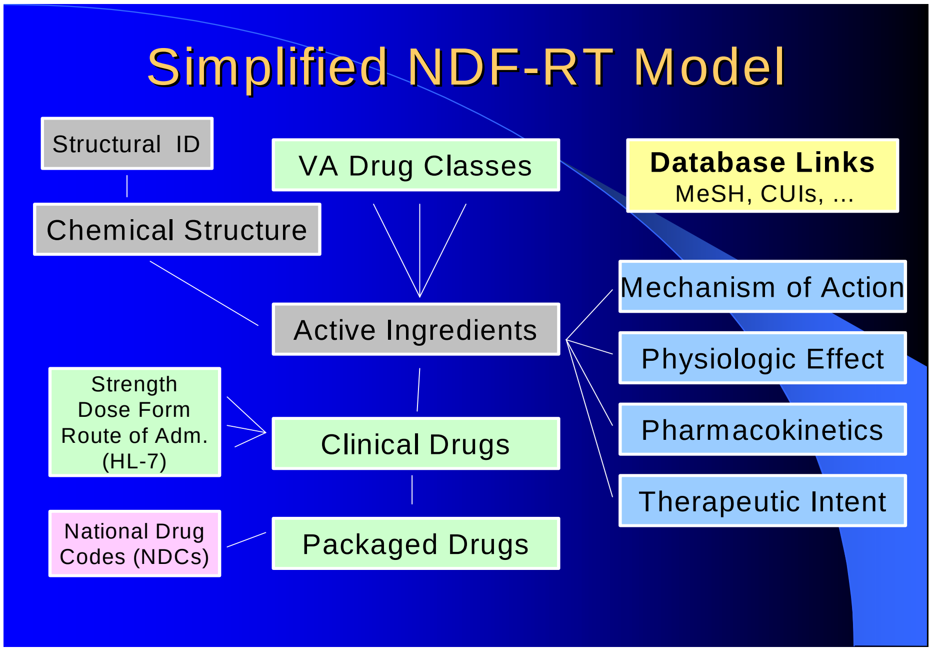
The current VHA formulary is the NDF. The NDF is:

- Centrally produced and maintained at VA.
- Locally modified and deployed.
- In use at more than 170 medical centers and thousands of clinics.
- Supporting automated fulfillment of more than 75 millions prescriptions per year.

Beginning with the NDF, the VHA is creating the NDF-RT (Reference Terminology) Model (Exhibit A-2), an ingredient-centric, semantic, i.e., computer-understandable, definition for each drug based on chemical structure, mechanism of action, physiologic effect, pharmacokinetics/dynamics, and therapeutic intent. The goal of the NDF-RT project is to evaluate the use of modern terminology techniques to increase functionality, improve quality and decrease costs.

The NDF-RT is a reference standard for medications to support a variety of clinical, administrative and analytical purposes. The area of clinical drugs is seen as important in the growing issues of patient safety. The NDF-RT Model is:

- Part of the VA Enterprise Reference Terminology (ERT) to rebuild the VA NDF.
- Linked to the VA Health Data Repository.
- An infrastructure for VA's CPRS reengineering.
- An explicit, multi-hierarchical, centered on ingredients.
- To develop new semi-automated terminology maintenance processes.
- Deployed via terminology server.
- To leverage authoritative, collaborative content.



Simplified NDF-RT Model

Structural ID

Chemical Structure

VA Drug Classes

Database Links
MeSH, CUIs, ...

Mechanism of Action

Active Ingredients

Physiologic Effect

Strength
Dose Form
Route of Adm.
(HL-7)

Clinical Drugs

Pharmacokinetics

Therapeutic Intent

National Drug
Codes (NDCs)

Packaged Drugs

**Exhibit A-2**. Simplified NDF-RT Model

The NDF-RT content includes:

- Drug Hierarchy
  - 3,977 Active Ingredients.
  - 11,345 Orderable Drugs.
  - 87,210 Packaged Drugs (NDCs).
  - Initialized from VA National Drug File (Sept 2001) and NLM RxNorm Drug –Names (December 2001).
- Reference Hierarchies
  - 3,994 Diseases and Manifestations ("Intended Therapeutic Use" hierarchy).

  - 489 Chemical Structure categories.
  - 402 Mechanism of Action and Physiologic Effect categories.
  - 154 HL7 Dose forms.
  - 58 Clinical Kinetics categories.
  - Initialized from MeSH and HL7.

Current status of NDF-RT:

- Model in place, auto-initialization complete and evaluated
- In progress.
  - PharmD content review complete, 2nd pass underway
  - Transactional maintenance model
  - Web-based update mechanism
  - Pharmacogenomic and cancer model extensions
- Exploring additional collaboration opportunities.
  - Explicit, multi-hierarchical model
  - Centered on drug ingredients for function, maintenance, and economies of scale

***Usage***

NDF-RT is a description logics enhanced drug reference terminology, which builds on the VA's NDF formulary drug terms. The FDA has been working in partnership with the National Library of Medicine (NLM) and the VA and in open collaboration within the HL7 standards development organization to promote and adopt the NDF-RT in the domains specified by the National Committee on Vital and Health Statistics (NCVHS).

NDF-RT is a recent addition to the UMLS. RxNorm, the representations of Mechanism of Action and Physiologic Effects for NDF-RT, and the proposed linked FDA terminologies have also been approved as U.S. government standards by the Consolidated Health Informatics (CHI) initiative.

***Applicability to NCI***

The NDF-RT is applicable to NCI because of our cooperative involvement in cancer drug data development and clinical trials. The NDF-RT model is particularly useful in drug-related clinical trials due to the fact that the model defines each drug based on chemical structure, mechanism of action, physiologic effect, pharmaco-kinetics/dynamics, and therapeutic intent. Therefore, NCI should consider adopting the NDF-RT and NCI Thesaurus models as a standard for its clinical trials.

***Curation***

NDF-RT is currently available on the NCI EVS Distributed Terminology Server and will be included in the first 2004 build of the NCI Metathesaurus.

***NCI Role***

The NCI will continue to work cooperatively with the VA on mutual enhancements and update processes for drug terminology in both NDF-RT and NCI Thesaurus. NDF-RT and NCI Thesaurus will be mapped in a 2004 release of the NCI Metathesaurus, and links provided in NCI Thesaurus to matching NDF-RT drug related concepts.

### 2.4.2 RxNorm Clinical Drug Vocabulary

<http://umlsinfo.nlm.nih.gov/RxNorm.html>

***Sponsor***

The National Library of Medicine (NLM), of the National Institutes of Health (NIH) in Bethesda, Maryland, is the world's largest medical library. The library collects materials in all areas of biomedicine and health care, as well as works on biomedical aspects of technology, the humanities, and the physical, life, and social sciences. The collections contain more than 6 million items - books, journals, technical reports, manuscripts, microfilms, photographs and images. NLM is a national resource for all U.S. health science libraries through a National Network of Libraries of Medicine.

In 1986, the NLM began a long-term research and development project to build a Unified Medical Language System (UMLS). The purpose of the UMLS is to aid the development of systems that help health professionals and researchers retrieve and integrate electronic biomedical information from a variety of sources and to make it easy for users to link disparate information systems, including computer-based patient records, bibliographic databases, factual databases, and expert systems. In late 2001, the NLM and the Department of Veterans Affairs (VA) began an experiment in modeling clinical drugs in the UMLS Metathesaurus. This project has become known as the RxNorm project.

***Description***

<http://www.nlm.nih.gov/mesh/semanticnorm.html>

The NDF-RT is an ingredient-centric model to describe clinical drugs, and the RxNorm model, as part of the UMLS Metathesaurus, is focused on improving interoperability of drug terminology. The two are complementary to each other.

The goal of the RxNorm is to define standard names for clinical drugs through:

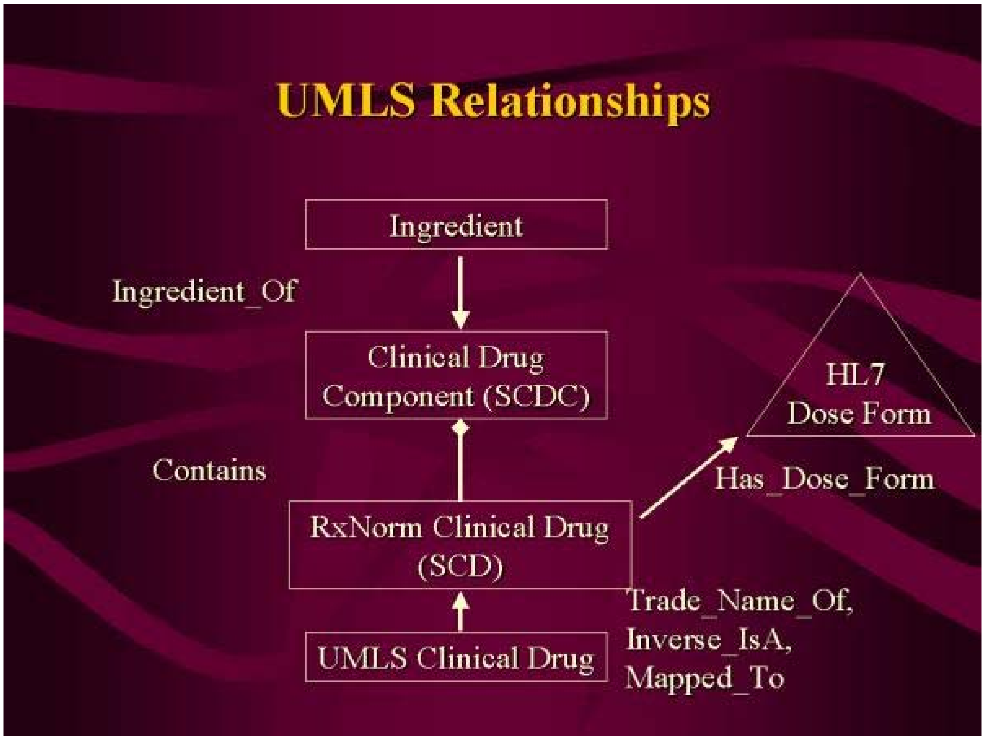
- Defining standard representational format
- Relating UMLS clinical drugs to the standard
- Facilitating navigation between vocabularies

The RxNorm project plan included:

- Establishing UMLS Concepts
  - Ingredients
  - Drug Components
    - › Ingredient and Strength
  - Dose Forms (from HL7)
  - Drug Formulations – The RxNorm Form
    - › Drug Component and Dose Form
- Standard Method for Representing Strength
- Relationships with Other Names

The RxNorm project approached clinical drug representation in a series of steps. The initial effort was to define a Semantic Normal Form (SNF) to represent clinical drugs. SNFs for clinical drugs are canonical representations, as defined by their active ingredients, strengths, and orderable dose forms. SNFs make explicit and/or normalize every active ingredient, strength, unit of measurement, and dosage form for a given clinical drug preparation. Clinical drug SNFs use standardized (generic) ingredient names, units, and dose forms, and a set of rules for expressing strength in a set of standard units.

There are two SNFs created as UMLS concepts for every clinical drug: the SNF Drug Component (SCDC) and the SNF Clinical Formulation (SCD). Exhibit A-3 depicts the UMLS clinical drug vocabulary model including the RxNorm form.



UMLS Relationships

Ingredient

Ingredient_Of

Clinical Drug Component (SCDC)

Contains

RxNorm Clinical Drug (SCD)

HL7 Dose Form

Has_Dose_Form

UMLS Clinical Drug

Trade_Name_Of,
Inverse_IsA,
Mapped_To

**Exhibit A-3**. UMLS Relationships

Preliminary results indicated that the model appeared to be adequate for expressing most of the orderable drugs. There are some areas that remain more problematic. Most multi-component ingredients fit into the model (though finding a suitable short name for generic multivitamins is a challenge), but others, such as additives for intravenous alimentation solutions, will require additional work. Other problematic areas are orderable materials used in tests (e.g., allergenic extracts), contrast media, and radiopharmaceuticals.

Vocabulary-specific route plus dose form combinations require mapping to the HL7 dose forms. Because each vocabulary is different in its expressions, this step must be done separately for each vocabulary. Similarly, ingredient names are not canonicalized or standardized. Since they are derived de novo from each candidate vocabulary, algorithmic determination of SNFs precise and base ingredient names will likely be imperfect or inconsistent.

This model for the SNFs of clinical drugs is intended to be useful for representing pharmaceuticals given to patients. It is possible that the model will be extensible to include such things as allergenic extracts, over-the-counter preparations, including herbal preparations and multivitamins, alimentation solutions, radioactive substances, and contrast media. However, it is not certain exactly how these will be approachable with this model. There will be further investigation required.

Additionally, there are devices containing drugs that may have more than one clinical drug in them (e.g., kits, oral contraceptive packs). Some of these cases may well be dealt with by establishing them as medical devices with a relationship attribute of "contains" to the SNF clinical drug.

***Usage***

Both NLM and VA started the RxNorm project as a research and development effort. The latest RxNorm Clinical Drug Vocabulary, a nonproprietary vocabulary that represents drugs at the level of granularity needed to support clinical practice, is UMLS Metathesaurus, which is distributed by the NLM quarterly as part of the UMLS project. NLM does not charge for the UMLS products. They are available to U.S. and international users. The full documentation for RxNorm is available through NLM. Nonetheless, the usage of the UMLS in commercial systems is low. In addition, it is not clear, how widely used the RxNorm Clinical Drug Vocabulary is among the current UMLS users.

***Applicability to NCI***

The RxNorm Clinical Drug Vocabulary is applicable to NCI because of its involvement in cancer drug development and clinical trials. As mentioned in previous section, the NDF-RT together with the RxNorm Clinical Drug Vocabulary can serve as standards to assist in the representation of developmental drugs. In addition, the adoption of the RxNorm and NDF-RT will help to achieve interagency efforts to promote and streamline the flow of information for drug data. Therefore, NCI should consider utilizing the RxNorm Clinical Drug Vocabulary.

***Curation***

RxNorm is included in the NCI Metathesaurus; hence it is available in the NCI EVS.

***NCI Role***

NCI and NLM have worked closely on a broad range of terminology. NCI is in a good position to influence the further development of the RxNorm Clinical Drug Vocabulary.

## 3. HEALTH-RELATED TRANSACTION STANDARDS

### 3.0.1 Digital Imaging and Communications in Medicine (DICOM)

<http://medical.nema.org/dicom/geninfo/dicom_strategy/index.html>

***Sponsor***

The DICOM initiative was created jointly by the American College of Radiology (ACR) and the National Electrical Manufacturers Association (NEMA). The DICOM standards organization is administered by the NEMA Diagnostic Imaging and Therapy Systems Division. Working groups of the DICOM Committee perform the majority of work on the extension of and corrections to the Standard in subject matter groups such as Cardiac and Vascular Information, Projection Radiography and Angiography, Nuclear Medicine, Compression, Exchange Media, Base Standard, Radiotherapy, Structured Reporting, Opthalmology, Strategic Advisory, Display

Function Standard, Ultrasound, Visible Light, Security, Digital Mammography and CAD, Magnetic Resonance, 3D, Integration of Imaging and Information Systems, Computed Tomography, and Dentistry.

***Description***

The DICOM Standards Committee exists to create and maintain international standards for communication of biomedical diagnostic and therapeutic information in disciplines that use digital images and associated data. The goals of DICOM are to achieve compatibility and to improve workflow efficiency between imaging systems and other information systems in healthcare environments worldwide. DICOM can be considered as a standard for communication across the “boundaries” between heterogeneous or disparate applications, devices and systems.

At the application layer, the services and information objects address five primary areas of functionality:

- Transmission and persistence of complete objects (such as images, waveforms and documents).
- Query and retrieval of such objects.
- Performance of specific actions (such as printing images on film).
- Workflow management (support of worklists and status information).
- Quality and consistency of image appearance (both for display and print).

When specific new technology is required, such as in support of new features such as security and compression, the strategy is to adopt proven international, industry or de facto standards. Accordingly, network confidentiality and peer authentication in DICOM are provided by the use of either TLS (an Internet standard) or ISCL (an ISO-based standard). Similarly, rather then develop medical-image-specific compression schemes, DICOM adopts standards developed by ISO/IEC JTC 1/SC 29/WG 1 such as JPEG and JPEG 2000. For interchange media, standard file systems compatible with conventional software (such as ISO 9660 and UDF) are used.

Version 3.0 of the Standard (released in 1993) specifies a network protocol utilizing TCP/IP, defined the operation of Service Classes beyond the simple transfer of data, and created a mechanism for uniquely identifying Information Objects as they are acted upon across the network. DICOM was also structured as a multi-part document in order to facilitate extension of the Standard. Additionally, DICOM defined Information Objects not only for images but also for patients, studies, reports, and other data groupings. The use of DICOM permits the transfer of medical images in a multi-vendor environment, and facilitates the development and expansion of picture archiving and communication systems (PACS) and interface with medical information systems

Following are links to the sections of the standard.

- DICOM Part 1: Introduction and Overview

- DICOM Part 2: Conformance

- DICOM Part 3: Information Object Definitions

- DICOM Part 4: Service Class Specifications

- DICOM Part 5: Data Structure and Semantics

- DICOM Part 6: Data Dictionary
- DICOM Part 7: Message Exchange
- DICOM Part 8: Network Communication Support for Message Exchange
- DICOM Part 10: Media Storage and File Format for Media Interchange
- DICOM Part 11: Media Storage Application Profiles
- DICOM Part 12: Media Formats and Physical Media for Media Interchange
- DICOM Part 14: Grayscale Standard Display Function
- DICOM Part 15: Security Profiles
- DICOM Part 16: Content Mapping Resource

### *Usage*

Every major diagnostic medical imaging vendor in the world has incorporated the DICOM standard into their product design and most are actively participating in the enhancement of the Standard.  DICOM is used or will soon be used by virtually every medical profession that utilizes images within the healthcare industry.  These include cardiology, dentistry, endoscopy, mammography, ophthalmology, orthopedics, pathology, pediatrics, radiation therapy, radiology, surgery, etc.  DICOM is even used in veterinary medical imaging applications.

DICOM has a number of optional components. Conformance to DICOM may apply to networks and/or to media storage. Not all users use all of the optional components, and any user can extend DICOM to meet specialized needs.

### *Applicability to NCI*

The NCI is a General Interest member of the DICOM standards committee. DICOM may be relevant to the identification of metadata associated with the storage of biomedical images.

### *Curation*

DICOM has a data dictionary that includes a listing of identifiers and names associated with data elements in the standard.  NCI will need to evaluate the DICOM standard to identify the components that are relevant to its programs.  It is possible that data elements from relevant component data sets could be registered in the caDSR, though additional metadata would need to be developed about each data element.

### *NCI Role*

NCI is already a member of the DICOM standards committee.

## 3.1 Basic Biology Transaction Standards

### 3.1.1 CellML$^{TM}$ Specification

<http://www.cellml.org/>

***Sponsor***

CellML is being developed by scientists at the Bioengineering Institute, The University of Auckland, New Zealand and Physiome Sciences, Inc.

***Description***

CellML is an XML-based language for describing and exchanging models of cellular and subcellular processes. As of 06/01/2004, CellML specification is in draft version 1.1. The development of CellML is guided by an advisory board drawn from many different areas of biological modeling and encourages reuse of models and parts of models by a component-based architecture. The components are logical parts that are connected together to form a model. CellML language is specifically designed for definition of model structure and is independent of a particular operating system or programming language. Mathematical information in CellML are incorporated using the Mathematical Markup Language (MathML, e.g. “<units name="fahrenheit"><unit multiplier="1.8" units="celsius" offset="32.0" /></units>” for temperature unit conversion), and metadata can be included using the Resource Description Framework (RDF, e.g. <rdf:Description rdf:about=""><!-- Some metadata content, such as a last-modified date --></rdf:Description>).

The <model> element is the root element of a CellML document, which could contain <import>, <units>, <component>, <group>, and <connection> elements that encode various components and their interactions in a cellular process. The CellML developers also provide software tools for editing, simulation environment, and XML validation.
CellML is intended for modeling all cellular processes, while the Systems Biology Markup Language (SBML) was initially developed with emphasis on modeling biochemical reactions and the interoperability of existing simulation tools. There are active collaborations between CellML and SBML developers and the model composition working group of SBML aims to ensure that the two representations can be integrated with each other to create a single standard in the future.

***License***

The official terms of use for CellML states:

> “Individuals can (a) freely use, publish, and redistribute the CellML Format and documentation; (b) write and sell applications which create, load, or write CellML-valid XML files; (c) distribute or sell their own CellML-valid XML files; and (d) transmit verbatim copies of the CellML Format and documentation to any person, without restriction. Applications and files using the CellML standard should reference http://www.cellml.org/public/specification/”

***Usage***

CellML can be used to describe models that simulate cellular processes. It is currently used by BioUML, Cellular Open Resource (COR), Cell Electrophysiology Simulation Environment

(CESE), Sheffield Epitheliome Project, LabHEART, University of New South Wales BME group, and Virtual Cell (VCell) Modeling and Simulation Framework.

***Applicability to NCI***

Similar to SBML, CellML could be a valuable tool that facilitates exchange of simulation models among broader NCI community. It is expected that CellML will eventually merge with SBML to become a unified standard for system biology simulation.

***NCI role***

None current.

### 3.1.2 The Systems Biology Markup Language (SBML)

<http://sbml.org/index.psp>

***Sponsor***

The SBML was suggested by Dr. Hamid Bolouri at the First Workshop on Software Platforms for Systems Biology in April 2000. The work was initially performed by Michael Hucka, Herbert Sauro, Andrew Finney and Hamid Bolouri at John Doyle's lab in the Control and Dynamical Systems Department (CDS) at the California Institute of Technology with support from the Japan Science And Technology Corporation's Exploratory Research for Advanced Technology program (JST ERATO). Currently, the SBML is very much a community effort.

***Description***

"The Systems Biology Markup Language (SBML) is a machine-readable format for describing qualitative and quantitative models of biochemical networks." The SBML is encoded in XML and designed to facilitate exchange of biochemical network models, with emphasis on quantitative models, between software tools. Current version of SBML is level 2, with level 3 in development.

"An SBML model consists of a set of chemical entities linked by reactions that can transform one entity into another or transport entities between compartments." The "Model" structure is the highest level construct that contains "FunctionDefinition", "UnitDefinition", "Compartment", "Species", "Parameter", "Rule", "Reaction" and "Event" components. The SBML level 2 uses the MathML standard by the W3 group to represent math elements and formulas, and the same metadata scheme as CellML. For example, one would represent the compartments of "cytosol" and "mitochondria" in a model using the following syntax:

```
<model>
        ...
                <listOfCompartments>
                        <compartment id="cytosol" size="2.5"/>
                        <compartment id="mitochondria" size="0.3"/>
                </listOfCompartments>
        ...
</model>
```

Currently, 44 software applications, including E-CELL, BioUML, PathArt, and SigPath, support the SBML standard according to SBML website.

***License***

The SBML is a free and open language.

***Usage***

System biology simulation models can be coded in SBML either locally or using one of the compliant tools, and shared between collaborating users, databases, or for publication purposes. This has a potential overlap with Cell ML. This is in use by a number of applications.

***Applicability to NCI***

Many basic mechanistic researches require quantitative simulation of cellular networks and systems. Using SBML to describe such models would facilitate sharing these models. It has potential applicability to the broader NCI community.

***NCI role***

None at this time.

## 3.2 Clinical Data Transaction Standards

### 3.2.1 Clinical Data Interchange Standards Consortium (CDISC) <http://www.cdisc.org/ >

***Sponsor***

CDISC is an open, multidisciplinary, nonprofit organization committed to developing industry standards to support the electronic acquisition, exchange, submission and archiving of clinical trials data and metadata for medical and biopharmaceutical product development. The mission of CDISC is to lead the development of global, vendor-neutral, platform-independent standards to improve data quality and accelerate product development in our industry.

CDISC works closely with the Food and Drug Administration (FDA) in eSubmissions and other areas of CDISC activities are also closely aligned with FDA activities, including laboratory data standards, standards to support data acquisition and archive, and metadata standards. NCI also works collaboratively with CDISC on controlled terminology assessment and development in support of clinical trials.

***Description***

CDISC, a global initiative, is leading the development of industry standards to support the electronic acquisition, exchange, submission, and archiving of clinical trials data and metadata for medical and biopharmaceutical product development. Exhibit A-4 depicts the scope of CDISC. Leading pharmaceutical, biotechnology, and information technology companies are uniting to support the CDISC.

CDISC supports a number of concurrent teams (and focus groups within these teams) that are working on specific topics. The teams include: the Operational Data Modeling (ODM) Team, the Submissions Data Standards (SDS) Team, the Testing and Applications Team, the Analysis Dataset Modeling (ADaM) Team, the Laboratory Data Team, and an Education Team. In addition, certain of these teams have External Review Committees to review data models in progress and provide formal feedback.

***Operational Data Modeling (ODM***). ODM refers to the standard data interchange models that are being developed to support the data acquisition, interchange, and archiving of operational data. The ODM is a vendor neutral, platform independent format for interchange and archive of data collected in clinical trials. The model represents general data (e.g., study name, protocol name, measurement unit), study metadata (e.g., code lists), administrative data, reference data (e.g., lab normal ranges), and clinical data associated with a clinical trial. Only the information that needs to be shared among different software systems during a trial, or archived after a trial, is included in the model. Systems that do not have all the features represented by the ODM model may still be ODM compatible as long as they comply with the conformity rules provided in the section on System Conformity.

Beginning with version 1.2, there is both an eXtensible Markup Language (XML) Schema and a Document Type Definition (DTD) provided with the ODM model. The XML schema is available. The DTD is informational and is provided for backward compatibility.

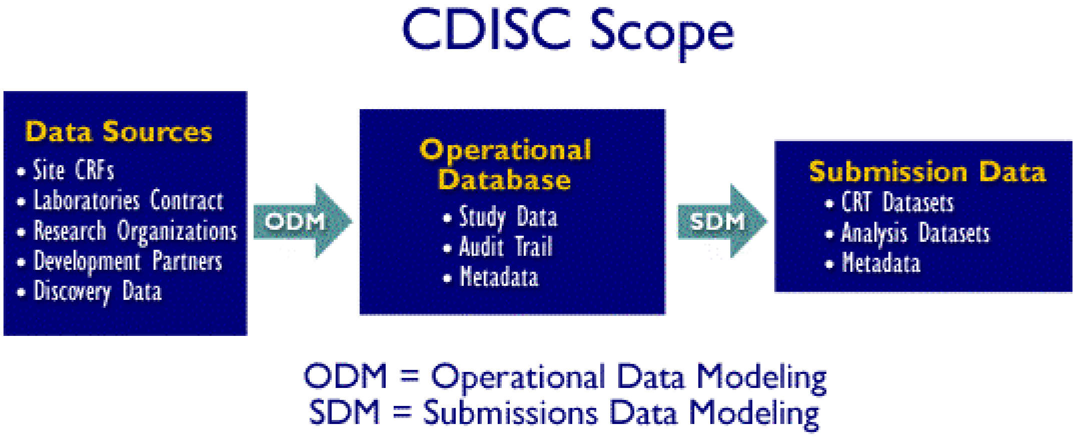


**Exhibit A-4**. CDISC Scope

ODM Version 1.1 Final (XML DTD for clinical data interchange and archive): The CDISC ODM Version 1.1 Final, released for implementation and comment in April 2002 significantly increases the usability and functionality of the ODM Version 1.0 DTD (released in October 2000). Version 1.1 Final replaces Version 1.1 Draft <http://www.cdisc.org/models/odm/v1.1/index.html> that was released in November 2001.

Submissions Data Standards (SDS). Submission Data Modeling (SDM) refers to the standard metadata models being developed to support the data flow from the operational database to regulatory submission. Version 3.0 (V3) of the CDISC Submission Data Domain Models has been prepared by the CDISC Submission Data Standards (SDS) Group to guide the organization, structure, and format of the Case Report Tabulation datasets for clinical trials submitted to the FDA. V3 is intended to facilitate transfer of data from sponsors to the FDA and subsequent loading into the FDA repository. V3 is not intended to meet the needs currently supported by analysis datasets, which will continue to be submitted separately. The CDISC ADaM team is developing analysis metadata models separately.

The CDISC Version 3 <http://www.cdisc.org/models/sds/v3.0/index.html> submission data standards have been approved as an HL7 informative document, and FDA reviewers and statisticians are currently piloting the standard with nine major global pharmaceutical companies. The resulting version of these standards (SDS Version 3.1) is scheduled to be directly referenced by FDA Guidance by mid 2004. The SDS Version 3.1 will simultaneously be balloted by HL7. CDISC is developing this model in conjunction with HL7, and that the HL7 balloted version will be constructed using the HL7 RIM.

Version 2.0, <http://www.cdisc.org/models/sds/v2.0/index.html>of the SDS, published in November 2001, incorporates several improvements to the Version 1.0 models published in October 2000 and the Version 1.1 revision published in June 2001.

Analysis Dataset Model (ADaM). The ADaM team has developed a guideline and several examples for analysis datasets used to generate the statistical results for a regulatory submission. The document "Guideline for the Creation of Analysis Files and Documentation of Statistical Analyses for Submission to the FDA" provides guidelines for creating analysis data sets and associated documentation that are submitted to the FDA statistical reviewer to support the primary and important secondary study objectives. These guidelines discuss the various issues that should be considered when submitting information to aid the statistical review of the submission. The guideline can be found at: <http://www.cdisc.org/models/adam/ADaM_Guidelines_V1.pdf>.

Laboratory Standards (LAB). The first version of the Lab Model, the CDISC Lab Model V1.0.0, was published in November 2002. The Lab Model has been developed through multiple iterations and tested with representative laboratory data. The Model was then further revised to address comments provided by a 65 member Laboratory Review Committee of industry experts and a 60-day public review. The Lab V1.0.1 release consists of the Base Model, the Schema Representation, and the Microbiology Extension. The CDISC LAB Base Model Version 1.0.1 is being released to industry for implementation. This update incorporates changes to model specifications to address minor bugs and to harmonize with CDISC's ODM and SDS models. Concurrent with this release, the CDISC LAB Team is posting a draft XML schema representation of the model for comment as well as a draft microbiology extension for comment. There are, in addition, sample datasets being provided to further illustrate the use of the Lab Model <http://www.cdisc.org/models/lab/v1.0.1/index.html>. The standard is now approved as an HL7 Reference Information Model (RIM) Version 3 message, in addition to the original ASCII, SAS and XML implementation options.

### *Usage*

Standards set by the CDISC for electronic data interchange hold the potential to speed clinical trials information management and make them less expensive and more accurate. The CDISC has a broad support of the leading pharmaceutical, biotechnology, and information technology companies (e.g., Amgen, Aventis, Merck and Co., Inc., Eli Lilly and Company, Novartis, Schering-Plough Corporation, SmithKline Beecham, DataTrak, Domain Pharma Corporation, Duke Clinical Research Institute, Fast Track Systems, NextPhase International, PHT Clinical Networks, and Quintiles Transnational). Also the European Union (EU) and Japan have supported and adopted the CDISC. In addition, the FDA considers CDISC as the most important partner to develop the standards, and they are committed to their adoption once the work is completed.  The resulting version of these standards (SDS Version 3.1) is scheduled to be directly referenced by FDA Guidance by mid 2004.  The SDS Version 3.1 will simultaneously be balloted by HL7.

***Applicability to NCI***

The CDISC is the first organization to simultaneously develop a set of standards to address electronic acquisition, exchange, submission and archiving of clinical trials data and metadata for medical and biopharmaceutical product development.  Based on NCI's mission and its close collaboration with the pharmaceutical industry, the CDISC standards are applicable to NCI's research data. The CDISC board has expressed a commitment to work jointly with the NCI, FDA, and HL7 organizations on the development and refinement of these common models, including a controlled terminology set for clinical trials.  Many CDISC standards may be balloted by HL7, which would result in their increased adoption.

***Curation***

CDISC standards are excellent candidates for registration of metadata in the caDSR, and representation of standardized terminology sets in EVS.  There should be a decision made regarding timing of inclusion because the standards are still evolving.

***NCI Role***

CDISC is open to all who want to participate.  One can participate by attending CDISC meetings, providing feedback on CDISC models through the Website Discussion facility, attending presentations at conferences, joining as a CDISC Corporate Sponsor or Corporate Member, working with support groups such as the Glossary Group, and/or participating in the working teams.  In addition, there are currently four options for membership in CDISC: Corporate Benefactor, Corporate Sponsor, Corporate Member and Associate Member. The Industry Advisory Board (IAB) comprised of one representative from each Benefactor or Sponsor company who provide advice to CDISC on strategic direction and support the strategic plan; participate in advisory board meetings, working teams, and task forces.  The IAB selects one member to sit on the CDISC Board of Directors each year.

NCI personnel are actively working with CDISC members through the HL7 RCRIM committee on these standards.  The collaboration of FDA, HL7, CDISC and other working partners through HL7 provides a mechanism for creating a single health-related transaction standard applicable to NCI.

### 3.2.2 Health Level Seven (HL7)
<**http://www.hl7.org/**>

***Sponsor***

Health Level Seven (HL7) is a developer of standards to support electronic communication between applications in the health care domain.  The HL7 organization was founded March 1987 in a meeting sponsored by the University of Pennsylvania. Over the intervening 16 years, there have been multiple versions of the messaging standard produced, and the scope of the organization has grown both functionally and geographically.

The original HL7 standard addressed the primary needs for application interoperability within a hospital setting. These needs most notably include transmission of patient and patient stay information, clinical orders, and clinical results. As interest and participation in the standard developed, this scope grew to include a wider range of data exchange within the hospital, as well as exchanges outside of the hospital setting.

HL7 had its beginning as a U.S.-based organization. Over time, its scope has become internationalized through the addition of affiliates based in numerous countries. Currently, there are 24 international affiliates across five continents.

### *Description*

The introduction to HL7 Version 2.5 states: "The Standard currently addresses the interfaces among various systems that send or receive patient admissions/registration, discharge or transfer (ADT) data, queries, resource and patient scheduling, orders, results, clinical observations, billing, master file update information, medical records, scheduling, patient referral, and patient care. It does not try to assume a particular architecture with respect to the placement of data within applications but is designed to support a central patient care system as well as a more distributed environment where data resides in departmental systems. Instead, HL7 serves as a way for inherently disparate applications and data architectures operating in a heterogeneous system environment to communicate with each other."[16]

In terms of this review effort, it is important to recognize that HL7 is both a data element standard and a vocabulary standard. That is to say, the messaging (transaction) specifications include a definition of standard data elements and a definition of the allowable vocabularies for coded elements.

It is important, when considering the role of HL7 standards, to clearly distinguish between the Version 2 and Version 3 families of HL7 standards. The Version 2 family includes the earliest HL7 standards and includes the implemented versions: 2.1, 2.2, 2.3, 2.3.1, 2.4, and 2.5. This list of standards includes the bulk of the HL7 implementations. However, the V2.x contents are not drawn from the HL7 reference models, and it is expected that, over time, implementations will migrate to the emerging V3 standards. HL7 Version 3, which represents a new departure for HL7, is still under development, and has been rolled out in only a few sites. The Version 3 specifications are constructed using an explicit message development methodology that is predicated on the use of common models of clinical data and terminology. This method is intended to lead to a more cohesive and useful set of standards.

Material related to HL7 is drawn from the balloted standard - across its several versions - and from the HL7 Website.

Health Level Seven Data Element Standards

HL7 is a messaging standard. At the highest level, it provides definitions of messages (transactions), and trigger events. The material below is relevant, because the data elements defined by HL7 can only be understood within this context. The standard notes: "The Standard is written from the assumption that an event in the real world of healthcare creates the need for data to flow among systems. The real-world event is called the trigger event. For example, the trigger event, a patient is admitted, may cause the need for data about that patient to be sent to a number

[16] Health Level Seven Messaging Standard Version 2.5, Chapter 1, Page 3.

of other systems. The trigger event, an observation (e.g., a CBC result) for a patient is available**,** may cause the need for that observation to be sent to a number of other systems."[17] In this context, the message specification defines the array of data elements that is transmitted when a particular trigger event occurs. This array of data elements is known as an "abstract message." An abstract message is essentially a hierarchical structure associated with a trigger event that defines the kind of data that is needed to support the trigger event.

A segment is a group of fields, each of which is defined by a particular datatype. Fields can have a simple or complex structure. They are constructed of components according to the rules defined in their datatype definition.

In the first HL7 specifications, there was no definition of the abstract message as such; it was simply the pattern of segments associated with a trigger event. Similarly, HL7 messages contain collections of segments that repeat together, segment groups.

The standard also includes text definitions for each field, some of which are extensive and revealing. The reader should note that the combination of Segment Mnemonic, e.g., Patient Identification Segment (PID), and sequence uniquely identifies each field within a segment. This combination is used within the XML encoding to indicate segments and is in much interface documentation. It is also important to note the datatype declaration for each field as well as the table number that applies to coded elements.

The datatype specification is an important tool for partitioning the complexity of the HL7 standard and is critical to understanding the data contents of an HL7 field. Some datatypes are simple and contain only one component; some contain many components and sub-components. For example, PID.5 Patient Name, has the datatype XPN in Version 2.4. This datatype supports the common subdivisions of an English language name, e.g., surname, first name, middle name, as well as suffix, prefix, name type code, and name validity (date) range.

Version 3 provides a new direction for HL7 messaging. It is intended to build on the experience of constructing and implementing Version 2. It is designed to provide a messaging standard that is built on a more solid foundation than its predecessor, and that will be easier to implement, easier to extend geographically, and introduces new functionality.

The driving force for V3 development was the interoperability issues that some V2 implementers have and the need to support messaging in a wide range of functional and national settings. Development of a new version of HL7 has first required the creation of a new methodology - a framework for message construction.

In HL7 Version 3, message specifications are created through a process of information model restriction and development. In fact, the creation of appropriate information models is the core of the V3 process. All messages draw their content from the HL7 Reference Information Model. However, that model, which is highly generic, has to be both restricted and elaborated on to specify the domain content for a message specification.[18] This process leads to the construction of a family of related models:

[17] Ibid, Chapter 2, Page 4.

[18] The HL7 Message Development Framework describes how classes from the RIM can be re-used and given specialized names, and how their contents can be restricted in order to support the development of specific message models.

- **Reference Information Model (RIM)**: The contents of the RIM comprise the entire scope of HL7, and it is a basic methodological principle that any attribute (data element) used in a V3 message must be drawn from the RIM. The RIM has a highly generic and abstract structure in order to represent the entire body of a healthcare message without being overly large and complex.
- **Domain Message Information Model (DMIM)**: A DMIM is a model covering an appropriate functional area within the scope of an individual HL7 Technical Committee. The DMIM is derived from the RIM through a process of sub-setting and extension (HL7 calls this latter aspect “cloning”). The end result is a model that contains classes, each of which is based on a class within the RIM, in which the attributes of each class are drawn directly from the RIM. A committee uses the DMIM as a development and expository tool.
- **Refined Message Information Model (RMIM):** The RMIM is a model of the content of a message or a closely related group of messages. RMIMs are developed to clearly specify the message's contents.

The normative products of HL7 V3 message development are message types drawn from a Hierarchical Message Description (HMD). The HMD is drawn directly from the RMIM; in fact, it is a "serialized" version of that model. That is to say that the HMD organizes the classes of the RMIM into a specific linear sequence to support the logical requirements of messaging.

Vocabulary Components of HL7

The importance of vocabulary is a key lesson that has been learned from health industry experience with interoperability through interfacing. Many interface partners have learned that agreement on message structure counts for little unless the vocabularies for coded elements are synchronized. HL7 Version 2 provides some support in this area, but for the most part, it provides a structure that transmits codes drawn from local coding systems.

In HL7 Version 2, all coded fields are linked to a table. The segment table includes the identifier of the table that is used for the field. There are three kinds of tables: HL7 defined, externally defined, and user defined. (The reader should note that, in some cases, the standard will supply example values for a user-defined table. These should be treated as they are labeled, as examples.)

An enhanced role for vocabulary management is a key feature of HL7’s development of Version 3. This enhanced role is based on the realization that sharing a common set of vocabularies is a key requirement for effective interoperability between systems. In Version 3, HL7 attempts to follow up this realization by a) creating a more systematic framework for managing vocabulary items, b) charging the HL7 Vocabulary Technical Committee with creating the proper linkages between Version 3 messages and externally developed and managed vocabularies, and c) creating a process that will allow vocabularies to be updated dynamically and independently of new standard releases of the standard. Accordingly, HL7 V3 has the goal of creating linkages between HL7 and widely accepted and used vocabularies, e.g., LOINC, SNOMED. HL7 also seeks to define principles for proper vocabulary development, and will itself - when no suitable industry vocabulary exists - define vocabularies.

The list of V3 vocabulary domains is included within the draft specification. However, the process of defining the set of codes within the individual domains is still in progress, and it would be premature to include that material here.

### *Usage*

HL7 is both widely used by health care providers and strongly promoted by U.S. government agencies. Among health care providers, it is clear that HL7 is, by far, the most prevalent health care interface standard. In fact, it has been said that 90 percent of U.S. acute care hospitals have implemented HL7.

In the government arena, Centers for Disease Control and Prevention (CDC) has long used HL7 for supporting vaccine registries, and for lab reporting related to notifiable disease. More recently, the use of industry standards, in particular HL7, is a prominent feature of the Public Health Information Network (PHIN) that CDC is developing. At the same time, the Food and Drug Administration (FDA) is working on HL7 specifications to address such areas as adverse event reporting, and drug stability reporting. HL7 is also working with the Clinical Data Interchange Standards Consortium (CDISC) to define standards for the interchange of information about, and generated by, regulated clinical research.

### *Applicability to NCI*

HL7 standards development covers a number of areas that are relevant to NCI. The following lists some of the principal points where HL7 work is relevant to NCI:

- HL7 has become the central organization for developing standards within the healthcare arena. The strategic role of HL7 is indicated by the wide range of participants in HL7 activities and is recognized in efforts by the U.S. government to improve patient safety and to enhance the cost effectiveness of healthcare delivery. It is noteworthy that HL7 has an active group working on issues related to clinical trials, and that work is ongoing in developing messages in this area. Much of the work that HL7 does in the vocabulary area - to develop vocabulary principles, in evaluating and coordinating with externally developed standards, in developing HL7 defined value sets, should be made available via the NCI EVS. It is also worth mentioning that HL7 has several groups whose work touches on the area of clinical protocols and would be relevant for the larger effort of achieving a common way to represent the data generated through cancer research.

- HL7 is an important developer of vocabulary items, and an even more important forum for defining the role of vocabularies such as SNOMED and LOINC in the context of clinical data interchange. It will be immediately useful to capture the value sets that have been defined for HL7 Version 2, and, as Version 3 matures, the value sets defined there should be captured as well. In order to make these items useful, it will be important to represent the context they are used in, and to provide descriptive information for potential users. These could be registered either in the caDSR as value domains or as terms in EVS.

- The data elements/attributes that are defined for use in HL7 messages are not particularly useful in isolation from their positioning within those messages. This is especially true of the elements and segments that have been defined for HL7 Version 2 messages. It would be possible to conceive of using HL7 messaging for communicating research data; however, outside of such use, there will not be great value in importing lists of data elements. This is particularly true because the value of the elements only emerges in the context of the data types used to define their properties.

- The development of a common data model - the Reference Information Model (RIM) - has been a key feature of the HL7 Version 3 process. This feature has led to the creation of a highly generic model that is used to enforce consistency and the use of common structures within the body of Version 3 data structures, e.g., messages and documents. This use of the RIM has great potential for assisting the NCI in its efforts to standardize across research projects; and consideration should be given to providing researchers and staff with further education on this model.

***Curation***

In this context, it is important to note that HL7 is working with NIST (National Institute of Standards and Technology) to develop a repository for HL7 messaging and vocabulary products. The contents of this repository would be available to HL7 members, and it has the potential to become the reliable source of HL7 material. NCI should track the development of this project because of its potential to ease the use of HL7 artifacts.

***NCI Role***

Membership in HL7 is available to everyone interested in the development of a cost-effective approach to system connectivity. NCI is both an Institutional member of HL7 and provides additional participation as well through the efforts of various NCI staff and contract personnel.

## 3.3 Genomics Transaction Standards

### 3.3.1 The Tissue MicroArray (TMA) data exchange specification

<http://www.biomedcentral.com/1472-6947/3/5>

***Sponsor***

The Association of Pathology Informatics (API) and the National Cancer Institute (NCI) co-sponsored a series of workshops to develop an open, community-supported TMA data exchange specification.

***Description***

Tissue microarray is a platform where hundreds of small tissue samples are mounted on a single glass slide, which allows the researchers to look at many tissues in one experimental sample. The TMA data exchange specification defines a standard for formatting all the data of a TMA experiment that is easy to be made understandable for both humans and computer programs.

> “The TMA data exchange specification is a well-formed XML document with four required sections: 1) Header, containing the specification Dublin Core identifiers, 2) Block, describing the paraffin-embedded array of tissues, 3)Slide, describing the glass slides produced from the Block, and 4) Core, containing all data related to the individual tissue samples contained in the array.”

As of October 2002, the TMA data exchange specification contains 80 Common Data Elements (CDEs, such as “slide section-thickness”, or “core_anatomic-site”) and 6 simple semantic rules (e.g. “2. Every TMA file must have histo as its root element.”). A perl script is also available for validating a TMA data file.  It is conceivable that some of the sample description vocabularies would overlap with similar terminologies in the histology and histochemistry.

### *License*

The TMA data exchange specification is open and freely available for the scientific community to use.

### *Usage*

The TMA data exchange specification is an open standard that once adopted can greatly facilitate exchange of TMA data between different data generators and databases. It can also been used as a guide for designing database schemas of TMA data repositories.

### *Applicability to NCI*

TMA is an important research platform for cancer research and NCI is a sponsor of the TMA data exchange specification.  An XML file representation will be potentially usable for the caBIG Tissue Banks and Pathology Tools workspace.  There are no competing standards, so it is a good candidate for use with microarray data.

### *NCI role*

NCI was involved in the development of the XML file representation.

# APPENDIX B – Standards for Potential NCI Consideration

## 1.0 COMMON DEMOGRAPHIC/INFORMATION PROCESSING AND CODE SETS

## 1.1 Address

### 1.1.2 United States Postal Service (USPS) Postal Addressing Standards

<http://pe.usps.gov/cpim/ftp/pubs/Pub28/pub28.pdf>

***Sponsor***

United States Postal Service (USPS)

***Description***

Jointly developed by the Postal Service and mailing industry, the Postal Addressing Standards (USPS Publication 28), include the uniform methods for matching addresses with the information in Address Information System (AIS) products and formats for outputting addresses on mail pieces. The standards describe both standardized address formats and content and outlines the guidelines that govern how address information appears in AIS products. Format describes how the various elements appear on a mail piece or in an address record. Content describes the characters that constitute the various address elements. The standard also defines business-to-business data elements for Company/Contact Information (e.g., Name Prefix, Surname, Professional Title) and Distribution and Delivery Address Information (e.g., Street Number, Company Name, ZIP+4 Code).

***Usage***

Postal units are adopting these standards, which are required by all internal processing systems and our licensees. This standard includes the National Change of Address (NCOA) System and Address Change Service (ACS). Mailers are encouraged to incorporate the standards as a means to improve service and deliverability.

***Applicability to NCI***

The standards would be applicable to systems requiring the exchange of mailing address data among users and internal data systems where such use contributes to operational benefits, efficiency, and economy. Furthermore, use of a single standard for representing addresses will make it easier to carry out locational analysis of the provided data.

***Curation***

It may be worthwhile to register this standard in the caDSR as a reference for anyone designing address data elements. These data elements would need to be clearly distinguished from the address-related data elements already registered in active contexts.

The addressing standard is described in USPS publication 28,
<http://pe.usps.gov/cpim/ftp/pubs/Pub28/pub28.pdf>

The USPS maintains several documents of USPS-recognized data element domains that could also be registered.

- USPS Official Abbreviations for States and Possessions.
- USPS Official Abbreviations for Street Suffixes.
- USPS Official Abbreviations for Secondary Unit Designators.

***NCI Role***

None

***Point of Contact***

United States Postal Services
International Postal Affairs
475 L'Enfant Plaza SW.
Rm. 370 IBU
Washington, DC 20260-6500
(202) 268-2444

# 2. HEALTH-RELATED VOCABULARY/CODING STANDARDS

## 2.1 Basic Biology Vocabulary Standards

### 2.1.1 Biological Pathways Exchange (BioPAX)

<http://www.biopax.org/>

***Sponsor***

The BioPAX project was initiated at the Fourth BioPathways Consortium Meeting, a satellite of the ISMB'02 Conference held in Edmonton, Canada in August 2002. It is funded with grants from the Department of Energy.

***Description***

The BioPAX group is developing a standard data exchange format for pathway information as a first step to build an open source pathway information resource and facilitate exchange of pathway information between existing databases.  BioPAX Level 1 is in the draft release version 0.5.2.

The BioPAX standard defines an ontology structure with "Entity" as the root class, and "Pathway", "Interaction", "physicalEntity" as the second level classes. Each 2nd level class contains sub-classes, such as "conversion" under "interaction", and "biochemicalReaction" under "conversion", and so on.

***License***

The BioPAX ontology files and accompanying documentation are freely available under the GNU LGPL license.

***Usage***

The BioPAX ontology is still in its early stages of development and we are not aware of any pathway database that has implemented the ontology.

***Applicability to NCI***

As this is under development, it is not ready for deployment at this time.

***NCI role***

NCI should monitor its development and consider its potential for a data exchange standard.

## 2.2 Clinical Vocabulary Vocabulary/Coding Standards

### 2.2.1 Current Procedural Terminology (CPT) 4

<http://www.ama-assn.org/ama/pub/category/3113.html>

***Sponsor***

The American Medical Association (AMA) was founded in 1847 with the following goals:

- Scientific advancement.
- Establishing standards for medical education.
- Launching a program of medical ethics.
- Improving public health.

First published in 1966 by the AMA, the Current Procedural Terminology (CPT) initially mainly focused on surgical procedures, with a limited number of other codes to describe medical, radiology, laboratory, and pathology procedures.

***Description***

The CPT is a comprehensive listing of medical terms and codes for the uniform coding of procedures and services that physicians provided. There were new editions published in 1970, 1973, and 1977. The fourth edition, CPT-4, contains more than 7000 new codes. Although there has not been a major revision of the CPT since 1977, CPT-4 is updated annually, with the newest version available each December.

CPT-4 contains a listing of all current U.S. FDA–approved physicians’ procedures and services. The AMA developed it in collaboration with various other health organizations. In the early 1980s, Congress decided to use CPT-4 to code all physicians’ procedures and services for Medicare patients. The aim of CPT-4 was to establish a way in which interested parties would know what procedures and services had been provided to the patient without reading a lengthy report.

CPT-4 is a system of 5-digit numeric codes and corresponding meanings. Every code is unique and is used only to describe a specific procedure, service, or medical supply physicians provided to their patients. This code holds true for inpatients and outpatients. Codes and descriptions are updated, revised, or changed yearly. The CPT-4 book is divided into six major sections:

1. Evaluation and management
2. Anesthesia
3. Surgery
4. Radiology
5. Pathology and laboratory
6. Medicine

Each section begins with its own specific guidelines and a listing of specific procedures and services applicable in that field. The guidelines contain definitions, explanatory notes, a listing of the previously unlisted procedures found in that particular section, directions on how to file a special report, modifiers for use in that particular section, and definitions to assist the coder.

Occasionally, a physician will perform a service that is not listed in the CPT-4 book. CPT provides unlisted codes at the beginning of each section for use when an unusual, variable, or new procedure is done. CPT-4 provides a way to give more information about a procedure through additional numbers called modifiers.

***Usage***

The CPT, a widely used mechanism in U.S. hospitals, clinics, and physician offices for grouping procedures and physician-provided services, is a proprietary coding scheme used primarily for billing. Additionally, the CPT-4 is one of the HIPAA-required data sets.

***Applicability to NCI***

Although the CPT-4 was created for billing purpose, it is still applicable to NCI. As part of a clinical trial, it is essential for investigators to adopt a standard coding scheme to capture any procedures and services provided by physicians to study participants during a clinical trial so that data can be recorded, analyzed, shared, and compared.

***Curation***

CPT-4 is included in the UMLS Metathesaurus as a Category 3 source. This means that use of CTP requires payment of a license fee to the AMA. CPT-4 is not currently included in the NCI Metathesaurus because no requirement for CPT-4 has been identified that justifies the license fee. If a requirement is established, CPT-4 can be added easily to the NCI Metathesaurus.

***NCI Role***

As mentioned, the CPT-4 is a proprietary coding scheme. Currently in order to establish new CPT codes, an individual, a physician, or a specialty group may submit a request to the AMA CPT Editorial Panel by using the Coding Change Request Form.

### 2.2.2 Healthcare Common Procedure Coding System (HCPCS)

<http://cms.hhs.gov/medicare/hcpcs/>

***Sponsor***

The Centers for Medicare and Medicaid Services (CMS) is a federal agency within the U.S. Department of Health and Human Services. Programs for which CMS is responsible include Medicare, Medicaid, State Children's Health Insurance Program (SCHIP), Health Insurance Portability and Accountability Act (HIPAA), and Clinical Laboratory Improvement Amendments (CLIA).

Because the AMA's CPT-4 codes do not include such items as ambulance service, wheelchairs, or injections, the CMS designed another coding system based on the CPT-4. This system is referred to as the Healthcare Common Procedure Coding System (HCPCS).

### *Description*

HCPCS is a medical code set that identifies health care procedures, equipment, and supplies for claim submission purposes. They are used in ambulatory settings. The HCPCS uses codes contained in CPT-4 (now known as HCPCS Level 1) plus expanded codes developed by CMS and fiscal intermediaries to classify physician and non-physician patient care services on the national level (now known as HCPCS Level 2). Level 2 HCPCS codes are most commonly referred to as the HCPCS codes, and Level 1 HCPCS codes are referred to as CPT. Since 1985, physicians have had to use the HCPCS to bill for services provided to Medicare patients either in the medical office or in the hospital.

Since October 1986, physicians also have used the HCPCS to bill for services provided to Medicaid patients. As of July 1, 1987, federal law requires hospitals to use the HCPCS to report outpatient surgery services to patients receiving health benefits sponsored by the federal government. By October 1, 1987, federal law had extended Ambulatory Surgical Center (ASC) prospective payment methodology to hospital outpatient surgery payments. The purpose of this was twofold: (1) To permit identification of ASC procedures so a blended payment rate could be applied to ambulatory surgery performed in the hospital outpatient department, and (2) To provide a database for future payment amounts for all hospital outpatient services.

The HCPCS includes three code levels that are discussed in the following sections.

### *HCPCS Level 1 Codes*

The HCPCS Level 1 codes are in CPT-4. This level lists terms and codes that provide a means to report physician procedures and services under both private and government-sponsored health insurance programs.

### *HCPCS Level 2 Codes*: National Codes

There is an HCPCS Level 2 code listing published once a year in the *National Coding Manual*, which can be ordered from the American Hospital Association, American Medical Association, or other publishers of the CPT coding book. It includes codes for the following:

- Chemotherapeutic drugs.
- Dental services.
- Durable medical equipment.
- Injections.
- Ophthalmology services.
- Orthotics.

- Some pathology and laboratory and rehabilitation supplies.
- Vision care.

National codes are 5-digit alphanumeric codes that begin with the letters A to V (e.g., L8100 is elastic support, elastic stocking, below knee, medium weight, each).

***HCPCS Level 3 Codes: Local Codes***

The HCPCS Level 3 codes were developed to address regional coding—the ability to code something performed or offered in one state that may or may not be performed or offered in another state. These codes, which are produced and made available through state Medicare carriers, may vary from state to state. Local codes begin with letters W to Z. CMS takes full responsibility for the codes in the *National Coding Manual*, leaving local codes up to Medicare carriers in each state.

***Usage***

Similar to the CPT-4, the HCPCS is a CMS-specific medical code set that identifies health care procedures, equipment, and supplies for claim submission to CMS for U.S. hospitals, clinics, and physician offices. Like the CPT-4, HCPCS is one of the required data sets by HIPAA.

***Applicability to NCI***

The HCPCS may not be applicable to NCI because the Level 1 codes are already included in the CPT-4, the Level 2 Codes are for classifying physician and non-physician patient care services, and the Level 3 Codes are regional-specific - not a national standard. It is unlikely that HCPCS codes are needed for the trial data collection and analysis, as well as for data sharing.

***Curation***

HCPCS is included in the NCI Metathesaurus.

***NCI Role***

If NCI wants to be involved in the further development of the HCPCS, NCI would need to work directly with the CMS.

### 2.2.3 International Classification of Diseases, Clinical Modification (ICD-9-CM). Ninth Revision

<http://www.cms.hhs.gov/medlearn/icd9code.asp>

***Sponsor***

The WHO is the United Nations specialized agency for health. WHO's objective, as set out in its constitution, is the attainment by all peoples of the highest possible level of health. The International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) is based on the WHO's International Classification of Diseases, Ninth Revision (ICD-9). It is a product of the Centers for Medicare and Medicaid Services, formerly the U.S. Health Care Finance Administration.

***Description***

Billing for patient encounters and other procedures must be accompanied by ICD-9 diagnosis codes. This coding system uses 3- to 5-digit codes to classify diseases, conditions, symptoms, complaints or problems by diagnosis. It was first initiated by an international panel but then modified by the United States National Center for Health Statistics (NCHS) on Clinical Classification. It has been the official system in the United States since 1977 for recording all diseases, injuries, impairment, symptoms, and causes of death.

The ICD-9-CM is used in assigning codes to diagnoses associated with inpatient, outpatient, and physician office utilization. It is the official system of assigning codes to diagnoses and procedures associated with hospital utilization in the United States. The ICD-9-CM classifies both diagnoses (Volumes 1 and 2) and procedures (Volume 3). Annual updates include the addition of new codes, removal of old ones, and revision of descriptors.

The ICD-9-CM consists of:

- A tabular listing of classification of diseases and injuries; supplementary classifications of factors influencing health status (V codes) and of external causes of injury and poisoning (E codes); and appendices for morphology of neoplasms, glossary of mental disorders, drug list numbers, industrial accidents, and 3-digit categories.
- An alphabetical index of diseases and injuries.
- A classification system for surgical, diagnostic, and therapeutic procedures (alphabetic index and tabular list).

The NCHS, the federal agency responsible for use of the ICD-10 in the United States, has developed a clinical modification of the classification for morbidity purposes. The ICD-10 is used to code and classify mortality data from death certificates, having replaced ICD-9 for this purpose as of January 1, 1999. ICD-10-CM is planned as the replacement for ICD-9-CM, volumes 1 and 2. Revisions have been made to the draft of ICD-10-CM based on the comments received. An updated draft version of ICD-10-CM <http://www.cdc.gov/nchs/about/otheract/icd9/icd10cm.htm> from June 2003 is currently available for public viewing. However, the codes in ICD-10-CM are not currently valid for any purpose or uses. Testing of ICD-10-CM will occur using this prerelease version. It is anticipated that updates to this draft will occur prior to ICD-10-CM implementation.

Notable improvements in the content and format of ICD-10-CM include:

- The addition of information relevant to ambulatory and managed care encounters.
- Expanded injury codes.
- The creation of combination diagnosis/symptom codes to reduce the number of codes needed to fully describe a condition.
- The addition of a 6th character.
- Incorporation of common 4- and 5-digit subclassifications.
- Laterality.
- Greater specificity in code assignment.
- The new structure will allow further expansion than was possible with ICD-9-CM. There is not yet an anticipated implementation date for the ICD-10-CM. Implementation will be based on the process for adopting standards under the HIPAA.

***Usage***

ICD-9-CM is the official system of assigning codes to diagnoses and procedures associated with hospital utilization in the United States. All hospitals and ambulatory care settings use this classification to capture diagnoses for administrative transactions. The ICD-9-CM procedure system is used for all inpatient procedure coding for administrative transactions. The ICD-9-CM Volume 3 Procedures were updated and distributed by the HHS. The ICD-9-CM has been selected under HIPAA as the required code set for use on the institutional claim format.

***Applicability to NCI***

ICD-9-CM was designed for use in billing and doesn't meet clinical research needs. It may have other applicability in the NCI, however.

***Curation***

ICD-9-CM is included in the NCI Metathesaurus that provides NCI researchers with access to ICD-9-CM. ICD9-CM is required for hospital billing. NCI will map to it, but is unlikely to use as a primary diagnosis terminology.

***NCI Role***

The NCHS and the CMS are the U.S. governmental agencies responsible for overseeing all changes and modifications to the ICD-9-CM. Recognizing the ICD-9-CM as a dynamic statistical tool that must be flexible to meet expanding classification needs, the ICD-9-CM Coordination and Maintenance Committee was created as a forum for proposals to update ICD-9-CM. A representative from the NCHS and one from the CMS cochair the ICD-9-CM Coordination and Maintenance Committee meetings. Responsibility for maintenance of the ICD-9-CM is divided between the two agencies, with classification of diagnoses (volumes 1 and 2) by NCHS and of procedures (volume 3) by CMS.

Although the ICD-9-CM Coordination and Maintenance Committee is a federal committee, suggestions for modifications come from both the public and private sectors. Interested parties are asked to submit recommendations for modification prior to a scheduled meeting. Proposals for a new code should include a description of the code being requested, and rationale for why the new code is needed. Supporting references and literature may also be submitted. Proposals should be consistent with the structure and conventions of the classification. NCI can use these processes to influence the further development of the ICD-9-CM and/or ICD-10-CM.

## 2.3 Genomics Vocabulary Standards

### 2.3.1 Mammalian Phenotype Ontology (MP)

<http://www.informatics.jax.org/searches/MP_form.shtml>

***Sponsor***

The MP is a developing standard initiated by the Mouse Genome Database (MGD) at the Jackson Laboratory since 1996. The development of MP is a community effort.

***Description***

As data from mutant animals accumulate from transgenic and gene knock-out experiments, there is a need for a standard vocabulary to describe and annotate these data. The MP specifies standard terms, such as “embryonic lethality”, or increased incidence of “pituitary adenoma”, to describe the phenotypes that may arise from certain mutations. Similar to Gene Ontology, these terms allows for uniform annotation of phenotypic data and allows query across different data sets and species for genetic alterations that result in similar phenotypes.

Unlike genomic and proteomic features, the domain of phenotype is highly diverse and specific to the species. The MP contains many phenotype descriptions and terminology that are equally applicable to humans as used by Online Mendelian Inheritance in Man (OMIM), although other parts (e.g. “tail”) are largely non-human.

***License***

The MP is part of the community developed Open Biological Ontologies (OBO) efforts, and is “provided to enhance knowledge and encourage progress in the scientific community and are to be used only for research and educational purposes.”

***Usage***

The MP terms should be used to describe any phenotype data, when these data are deposited into databases or exchanged between collaborating investigators. The MP terms can be downloaded from the Mouse Genome Informatics website <http://www.informatics.jax.org/searches/MP_form.shtml> at the Jackson Laboratory.
It is currently used and being developed by the Mouse Genome Database (MGD) at the Jackson Laboratory.

***Applicability to NCI***

The MP should be considered for inclusion into NCI vocabularies, since phenotypic descriptions of many mammalian studies are very much part of NCI research activities.

***NCI role***

The MP is still in a development stage and NCI should monitor the effort.

# 3. HEALTH-RELATED TRANSACTION STANDARDS

## 3.1 GENOMICS TRANSACTION STANDARDS

### 3.1.1 Macromolecular Structure (Mms)

<http://www.omg.org/technology/documents/formal/macro_molecular.htm>

***Sponsor***

The Macromolecular Structure (Mms) specification was submitted to the OMG by the National Institute of Standards and Technology (NIST). The OMG also acknowledges the contribution to the specification by the Research Collaboratory for Structural Bioinformatics (RCSB), and the San Diego Supercomputer Center. RCSB developed and currently operates the Protein Data Bank (PDB).

### *Description*

The Mms specification is a rather detailed data model definition for information related to macromolecular structures. The field of structural biology is one of the oldest sub-disciplines in biology, and one of the first areas in biological research to use computers for data management and as research tools. One of the benefits of the long history is that there is considerable accumulation of knowledge, such as the dictionary of terms developed by the International Union of Crystallography (IUCr) and the macromolecular Crystallographic Information File (mmCIF) standard by PDB. The Mms tries to adhere to the core IUCr dictionary, and its object oriented design allows for extensions of the core dictionary through sub-classes and inheritance.

The Mms contains a required MacromolecularStructure module and an optional MmsReference module. The MacromolecularStructure module contains elements describing related sequences to a structure, the atomic properties, chemical components, chemical bonds, secondary structures, active sites/domains, interaction between local structures, interaction between subunits of macromolecular complexes, etc. The MmsReference module defines elements that capture reference information on macromolecular structures, such as publication citations, computational methods and software, and database (e.g. PDB) related information.

### *License*

The Macromolecular Structure specification is an open standard and can be downloaded from the OMG website.

### *Usage*

The Macromolecular Structure specification defines both the data model for macromolecular structure, and adopts the consensus IUCr terms for data element naming and attribute values. Such specifications can be used for designing database schema to store structural data, or develop data interfaces for exchange structural data between different organizations.

PDB has developed the Mms standard and software, OpenMMS (http://openmms.sdsc.edu/), to use the standard. "The OpenMMS Toolkit contains software for parsing protein and nucleic acid macromolecular structure data stored in the standard mmCIF format. The toolkit also contains software for loading the mmCIF data files into a relational database, and for running a Corba server to deliver binary MMS data directly over a network connection to applications."

### *Applicability to NCI*

NCI does not have substantial macromolecular structure research and database activities, therefore the Mms standard does not appear to be applicable at this time. However, software applications that use macromolecular structural information should be developed using the Mms standard and OpenMMS toolkit.

Although NCI does not have substantial macromolecular structure research and database activities, it is not uncommon for molecular biologists to lookup 3-D structures of bio-molecules. The MMS is sponsored by the RCSB, that created PDB (the Genbank/EMBL/DDBJ equivalent for 3-D structures); and is of high quality and authoritative. Deployment could involve the addition of the OpenMMS JAVA source code into caBIO or caBIG or NCI could just host the MMS related services.  It may be beneficial to make the OpenMMS classes available for programming over caBIO for example.

*NCI role*

None.

### 3.1.2 PEDRo (Proteomics Experiment Data Repository) <http://pedro.man.ac.uk/>

*Sponsor*

PEDRo was developed by Norman Paton, Kevin Garwood, and Chris Taylor from the Department of Computer Science at the University of Manchester, and the European Bioinformatics Institute.

*Description*

PEDRo is a data model and associated software tool created "to support the capture, storage and dissemination of proteomics experimental data – necessary components for the implementation of a proteome repository." The PEDRo data model contains sections for storing information for sample generation (organism, tissue, cell type …), sample processing (fractionation, gel separation …), mass spectrometry (ion source, ion trap, Matrix-Assisted Laser Desorption/Ionization or MALDI …), and protein identification (peak, protein sequence database, database searching parameters …). The PEDRo data model has been implemented in both a relational database schema (SQL) and XML in the form of the Proteomics Experiment Markup Language (PEML), which is used in software tools developed by PEDRo.

The PEDRo standard has often been referenced by other data model and tool developers such as the SysBio-OM (SAIC-NIEHS) and SAS, although it may be superseded in the future by the developing standard by the Human Proteomic Organization (HUPO)'s Proteomic Standard Initiative (PSI). Although the future of the PEDRo standard is uncertain, some part of it may well be integrated into emerging standards developed by HUPO's Proteomics Standards Initiative as part of on-going collaboration.

*License*

All PEDRo standards and software releases are under the open source Academic Free License and can be downloaded from Sourceforge <http://sourceforge.net/projects/pedro>.

*Usage*

The PEDRo data model can be used as a guide for database developers to generate a relational database schema for mass spectrometry data. The PEML can be used for data exchange between software applications and databases or XML data files. In addition, the PEDRo developers have created software tools that can manage proteomics data using PEML and XML data files.

Although the PEDRo standard has not been officially endorsed by large proteomic consortiums, such as the HUPO, some organizations have started to use PEDRo due largely to lack of a proteomic standard. Examples of such organizations are the Consortium for the Functional Genomics of Microbial Eukaryotes (COGEME) and Medical College of Wisconsin Proteomics Center.

***Applicability to NCI***

As proteomic technologies are being adopted in cancer biology research, some sort of proteomics standards will be essential for developing informatics tools. In the absence of more mature standards, PEDRo should be used as a starting point with the recognition that such standards are evolving and changes are inevitable.

The PEDRo standard is in use at NIEHS for 2D Gel and Mass Spec. It is affiliated with the HUPO initiative. NCI hasn't used PEDRo yet, but there is a plan to reuse the NIEHS protein expression object model and portal to capture protein data in the Rembrandt/I-SPY project.

***NCI role***

NCI should monitor the use of PEDRo and assess its applicability to emerging NCI programs.

### 3.1.3 Protein-Protein Interaction (PPI)

<http://psidev.sourceforge.net/#proteinProteinInteraction>

***Sponsor***

The Protein-Protein Interaction standard is being developed by the Proteomics Standards Initiative (PSI) founded in 2002 by the Human Proteome Organization (HUPO).

***Description***

The PPI standard is being developed as a common data standard to bridge the different formats from the different protein interaction databases, such as BIND, DIP, MINT, Hybrigenics, and the MIPS. "The standard defines a minimal data model that allows scientists to provide core data, but refer back to the original data source for full information, in particular for complex, fully curated entries." The standard (version 1.0 as of 06/05/04) is formatted in XML, as PSI molecular interaction (MI) data exchange format designed by a group of people that include representatives from database providers and users.

The root element of a PSI MI XML file is the entrySet, which contains one or more entries. The top level elements in each entry, which describes one or more protein interactions, are "source" (source of the entry, e.g. BIND), "availabilityList" (availability of data, usually copyright statements), "experimentList" (experiment descriptions), "interactorList" (proteins that interact), "interactionList" (interaction elements), and "attributeList" (place holders for additional data type). Each interaction element contains information on the name the interaction, availability of the data, experimental evidence, type of interaction, protein participants, etc.

The PSI MI attempts to use externally developed controlled vocabularies where possible. Currently, the PSI provides controlled vocabularies for all the terms that are suggested to be used for the following areas:

- interaction type
- sequence feature type
- feature detection
- participant detection
- interaction detection

### *License*

The PSI MI is an open community developed standard that is available from the HUPO PSI website.

### *Usage*

The PSI MI is a data exchange format to facilitate data exchanges between users and databases (data submission and curation), and peer-to-peer (between users or between databases). It is envisioned by the PSI, that once the standards are adopted, protein interaction databases, such as BIND, DIP, MINT, Hybrigenics, and the MIPS, would be able to synchronize their data, similar to that of nucleotide data between EMBL, GenBank and DDBJ.

Although not originally intended, the PSI MI structure can also guide the design of protein interaction databases by providing some defined data elements and their relationships.

### *Applicability to NCI*

As molecular interactions are very important aspects of cancer biology, as well as biology in general, managing molecular interaction data is an essential part of NCI informatics support. The PSI MI provides an infrastructure for communicating with public protein-protein interaction databases and should be adopted, or harmonized with existing NCI data model.  It is premature to deploy this standard until the protein standards become more stable and a pattern of use is established.

### *NCI role*

NCI should monitor this standard.

# APPENDIX C
# Other Reviewed Standards that are Unlikely to Meet NCI Needs

## 1. Common Demographic/Information Processing and Code sets

## 1.1 Address

### 1.1.1 Federal Geographic Data Committee (FGDC) Address Data Content Standard
<http://www.fgdc.gov/>

***Sponsor***

Federal Geographic Data Committee (FGDC)

***Description***

The objective of the Draft Address Data Content Standard is to provide a method for documenting the content of address information for a physical location. As a data usability standard, the standard describes a way to express the content, applicability, and data quality and accuracy of a dataset or data element.

***Usage***

The standard establishes the requirements for documenting the content of addresses. It is applicable to addresses of entities having a spatial component. The standard does not apply to addresses of entities lacking a spatial component and specifically excludes electronic addresses, such as e-mail addresses. The standard additionally codifies some commonly used discrete units of address information, referred to as descriptive elements. It provides standardized terminology and definitions to alleviate inconsistencies in the use of descriptive elements and to simplify the documentation process.

***Applicability to NCI***

This address standard relates to geospatial data that would be encoded according to the FGDC data standards. It is unlikely to be applicable to most NCI data.

***Curation***

The Address Data Content Standard is available from the FGDC Web site. However, at this time, the standard is only draft. This geospatial address standard could also be confused with mailing address standard, which is more applicable to NCI business.

***NCI Role - None***

***Point of Contact***
Federal Geographic Data Committee
c/o U.S. Geological Survey
590 National Center
Reston, Virginia 22092

(703) 648-5514
gdc@usgs.gov

## 1.2 Vital Statistics

### 1.2.1 Bureau of the Census Current Population Survey

<http://www.census.gov/> <http://www.bls.census.gov/cps/cpsmain.htm>

***Sponsor***

The Current Population Survey (CPS) is a joint project between the Bureau of Labor Statistics and the Bureau of the Census.

***Description***

The CPS is the primary source of information on the labor force characteristics of the U.S. population. The sample is scientifically selected to represent the civilian noninstitutional population.

The survey has over 300 variables (depending on year). Household and family variables include type of living arrangement, structure, size of unit, income, earnings, health, and locational and date items related to survey taken. Personal variables include for labor force data employment (farm and nonfarm workers, persons self-employed, unpaid workers, wage and salaried employees), occupation of worker and industry of employment, number of hours worked, major activity last week, and reason for not working. Variables for demographic data include age, sex, race, ethnicity, Spanish origin, marital and family status, household relationship, children, veteran status, years of school completed, and place of residence. Supplementary data include migration, after-tax money income and the value of non-cash benefits (food stamps, school lunch programs, employer-provided group health insurance plans, employer-provided pension plans, personal health insurance, Medicaid, Medicare, CHAMPUS or military health care and energy assistance). Data on employment and income refer to the preceding year, and demographic data refer to the time of the survey.

***Usage***

The main purpose of the survey is to collect information on the employment situation, a very important secondary purpose is to collect information on demographic characteristics such as age, sex, race, marital status, educational attainment, family relationship, occupation, and industry.

***Applicability to NCI***

This survey includes a wide variety of commonly collected demographic variables, which may be of interest to NCI. However, it has a data coding scheme that is specific to the Bureau of the Census, and may be less useful outside the organization. It is not a true standard as it has not been reviewed by a variety of potential user organizations.

Some of the CPS codes and values that are of general interest include:

| **Category** | **Values** |
|---|---|
| Country | Proprietary two to three digit numbering scheme. |
| Age | 99 = 99 years or older<br>00-98 = 0 to 98 years old |
| Gender | 1 = Male<br>2 = Female |
| Marital Status | 1 = Married - spouse PRESENT<br>2 = Married - spouse ABSENT<br>3 = Widowed<br>4 = Divorced<br>5 = Separated<br>6 = Never married |
| Relationship of Family Members to Subject | 20 = Spouse (Husband/Wife)<br>21 = Unmarried Partner<br>22 = Child<br>23 = Grandchild<br>24 = Parent (Mother/Father)<br>25 = Brother/Sister<br>26 = Other relative (Aunt, Cousin, Nephew, Mother-in-law, etc.)<br>27 = Foster child<br>28 = Housemate/Roommate<br>29 = Roomer/Boarder<br>30 = Other nonrelative |
| Education Level | 31 = Less than 1st grade<br>32 = 1st, 2nd, 3rd, or 4th grade<br>33 = 5th or 6th grade<br>34 = 7th or 8th grade<br>35 = 9th grade<br>36 = 10th grade<br>37 = 11th grade<br>38 = 12th grade No Diploma<br>39 = High school graduate (Diploma, GED, or Equivalent)<br>40 = Some college but no degree<br>41 = Associate degree in college (Occupational/vocational program)<br>42 = Associate degree in college (Academic program)<br>43 = Bachelor’s degree (BA, AB, BS)<br>44 = Master’s degree (MA, MS, MEng, MEd, MSW, MBA)<br>45 = Professional school degree (MD, DDS, DVM, LLB, JD)<br>46 = Doctorate degree (PhD, EdD) |

**Exhibit C-1**. CPS Codes and Values Applicable to NCI

### *Curation*

The CPS Data Dictionary is available from the Current Population Survey Web site. While it may be of interest to illustrate how the Census Bureau structures demographic data elements, it is not recommended for registration in the caDSR or inclusion in the EVS as it is redundant with other standards which may be more applicable to NCI data.

***NCI Role***

None

***Points of Contact***

U.S. Census Bureau
4700 Silver Hill Road
Washington, DC 20233-0001
cpshelp@info.census.gov

<http://stats.bls.gov/>
U.S. Department of Labor
Bureau of Labor Statistics
Postal Square Building
2 Massachusetts Ave., NE
Washington, DC 20212-0001
(202) 691-5200
blsdata_staff@bls.gov

## 1.3 Education-related Standards

### 1.3.1. United Nations Educational, Scientific, and Cultural Organization (UNESCO) International Standard Classification of Education (ISCED)

<http://www.unesco.org/education/information/nfsunesco/doc/isced_1997.htm>

***Sponsor***

The International Standard Classification of Education (ISCED) was designed by United Nations Educational, Scientific, and Cultural Organization (UNESCO) in the early 1970's to serve as an instrument suitable for assembling, compiling, and presenting statistics of education both within individual countries and internationally.

***Description***

ISCED is designed to serve as an instrument suitable for assembling, compiling and presenting comparable indicators and statistics of education both within individual countries and internationally. ISCED has been designed to be universally valid and invariant to the particular circumstances of a national education system. The standard concepts, definitions and classifications cover all organized and sustained learning opportunities for children, youth and adults including those with special needs education, irrespective of the institution or entity providing them or the form in which they are delivered. The standard defines the concept of levels of learning experiences and the competences found in an education program that would provide a participant a reasonable expectation of acquiring the knowledge, skills and capabilities.

***Applicability to NCI***

It is not clear that this standard applies to NCI. A standard for the classification of educational levels of providers might be useful, but it does not appear that this standard would meet that need as the ISCED levels of education require quite a bit of interpretation to apply. It is designed for classification of educational programs, not individuals.

***Usage***

Not applicable at this time.

***NCI Role***

No participation in this standard is needed.

***Point of Contact***

UNESCO
7, Place de Fontenoy
75352 PARIS 07 SP, France
bpiweb@unesco.org

### 1.3.2 National Center for Education Statistics (NCES) <http://nces.ed.gov/>

***Sponsor***

The U.S. Department of Education has been involved in numerous efforts focused on improving the quality and comparability of educational data collected at the local, state, and national levels. NCES has taken the lead in many of these efforts and has a mandate to collect uniform and comparable data reporting the condition of education in the United States.

***Description***

NCES, working with federal, state, and local education agencies and researchers, has developed a series of data handbooks to provide guidance on consistency in data definitions for education data. These handbooks include data elements, definitions, and valid values for the following domains: School, Student, Staff, Local Education Agency, Intermediate Educational Unit and State Education Agency. Data elements are defined for areas including Person Demographics (e.g., Name, Sex, Address, Telephone Numbers, Electronic Mail Addresses, and Related Organizations), Health Records, Student Test Monitoring, Personnel Issues, Educational Achievement, School and Curriculum Management, and Assessments and Regulations.

***Applicability to NCI***

The NCES work can provide insight into the data definitions and formats used in data collection instruments in the educational field, particularly in the domains of Person Demographics and Student Health Records. The value domain information provided in the Appendices could provide guidance for determining code sets and field sizes for NCI information. While it may be a useful reference for review during NCI's harmonization, it is not a standard.

### *Curation*

Handbook information is available for use either as PDF or XML files containing data element names, definitions, and valid values. Registration of the set of NCES data elements in the caDSR is not recommended, as they do not constitute a relevant standard for potential adoption by NCI.

### *NCI Role*

Handbook information would be used as reference material only. As this information is specific to data collection for educational statistics, it is unlikely that NCI would be interested in participation.

### *Point of Contact*

National Center for Education Statistics
Institute of Education Sciences
U.S. Department of Education
1990 K Street, NW.
Washington, DC 20006

## 2. VOCABULARY/CODING STANDARDS

## 2.1 Mouse Pathology (MPATH)

< http://eulep.anat.cam.ac.uk/>

### *Sponsor*

The pathology ontology is part of the mutant mouse pathology database, or Pathbase that "is funded under the European Commission's Fifth framework programme (QLRI-CT-1999-00320) as a thematic network with the aim of providing a community resource for transgenic and mutant rodent pathology. The responsible Officer is Dr. Angeles Rodriguez-Pena." The co-ordinator of the project is the Dept. of Anatomy, University of Cambridge, UK; while the partners and subcontractors are: School of Biology and Biochemistry, University of Bath, UK; Dept of Pathology B35, University Hospital of Liege, Belgium; Department of Pathology, Swedish University of Agricultural Sciences, Sweden; ENEA, Divisione Protezione dell'uomo e degli Ecosistemi, Italy; Paterson Institute, The Christie Hospital, UK; The Research Institute, Churchill hospital, University of Oxford, UK; Radiation Sciences Centre, Dublin Institute of Technology, Ireland; Institut fuer Pathologie, GSF-Forschungszentrum fuer Umwelt und Gesundheit Ingolstaedter Landstrasse 1, Germany; Department of Biomedical Sciences Hugh Robson Building, George Square, Edinburgh University, UK; Institute of Biomedicine, Developmental Biology, University of Helsinki, Finland; Karolinska InstituteUnit for Morphological phenotype Analysis, Clinical Research Centre, Huddinge Hospital, Sweden; National Radiological Protection Board, UK.

### *Description*

The Pathbase pathology ontology is a structured vocabulary developed as part of the Pathbase. The Pathbase "is a database of histopathology photomicrographs and macroscopic images derived from mutant or genetically manipulated mice". Lesions are described using pathology terms organized in a hierarchical structure.

The pathology ontology overlaps with many medically derived terminologies, such as SNOMED and NCI Thesaurus. But the collection of pathology/histopathology images in Pathbase which is annotated with the MPATH ontology represents a valuable resource for pathologists.

***License***

The MPATH ontology is part of the community supported OBO projects, and is “open and can be used by all without any constraint other than that their origin must be acknowledged and they cannot be altered and redistributed under the same name.”

***Usage***

The MPATH ontology terms are used by the Pathbase to annotate the histopathology images. The pathology ontology terms can also be used to annotate microarray or proteomic experiments and samples, where pathological conditions are present.

***Applicability to NCI***

NCI will need standard vocabulary for histopathology. However, other standards used by NCI, such as NCI Thesaurus, ICD, and other terminology collections in the NCI Metathesaurus may already provide terms that cover the knowledge domain. Some efforts may be needed to incorporate MPATH into existing vocabularies, e.g. to take advantage of the Pathbase.

***NCI role***

None at this time.

## 2.2 Sequence Ontology Project (SO)

< http://song.sourceforge.net/>

***Sponsor***

The SO project is a “joint effort by genome annotation centres, including: WormBase, the Berkeley Drosophila Genome Project, FlyBase, the Mouse Genome Informatics group, and the Sanger Institute”. “We are a part of the Gene Ontology Project and our aim is to develop an ontology suitable for describing biological sequences.”

***Description***

The following quote is from the SO website which gives a very concise description of the project.

> “The Sequence Ontology is a set of terms used to describe features on a nucleotide or protein sequence. It encompasses both "raw" features, such as nucleotide similarity hits, and interpretations, such as gene models. It is also intended to have a rich set of attributes, such as "dicistronic" gene and "pseudogene."

The Sequence Ontologies are provided as a resource for the bioinformatics community. They have the following obvious uses:

- To provide for a structured controlled vocabulary for the description of primary annotations of nucleic acid sequence, e.g. the annotations shared by a DAS server.

- To provide for a structured representation of these annotations within genomic databases. Were genes within model organism databases to be annotated with these terms then it would be possible to query all these databases for, for example, all genes whose transcripts are edited, or trans-spliced, or are bound by a particular protein.

- To provide a structured controlled vocabulary for the description of mutations at both sequence and more gross level in the context of genomic databases. We have also defined attributes for many of the terms. Pro tem these are held in the "comment:" field of the definitions file.

> The four major nodes of the complete Sequence Ontology (so.ontology) allow for the annotation of chromosome mutations; the consequences of mutation objects that can be located (in base coordinates) on a sequence and general attributes of genes. A "lite" version of SO is also available (sofa.ontology) which includes only locatable features and is designed for use in such outputs as GFF."

The SO terms can be incorporated into bioinformatics tools for genome and sequence analysis based on the Biomolecular Sequence Analysis (BSA) data model specification.  There are considerable overlap between SO and the NCBI data model which is encoded in either Abstract Syntax Notation One (ASN.1) or XML format. However, SO is specifically focused on genome annotations and jointly developed by WormBase, the Berkeley Drosophila Genome Project, FlyBase, the Mouse Genome Informatics group, and the Sanger Institute, which could see its quick adoption by the genome databases.

### *License*

The SO is part of the community supported OBO projects, and is "open and can be used by all without any constraint other than that their origin must be acknowledged and they cannot be altered and redistributed under the same name."

### *Usage*

The SO terms are be used for annotating biological sequences, such as DNA, RNA, and protein. These terms should be incorporated into sequence submission, annotation, and curation tools. The SO is still maturing and few genomic databases have indicated usage of SO for genomic annotation. However, it is expected that the sponsoring databases would adopt the SO terms for new sequence entries and annotate their legacy data as well.

The SO represents an evolution of sequence data management in that there are large bodies of sequence data stored in major genome and protein databases, such as Genbank and EMBL, prior to the existence of the standards. To gain full benefits of the data standards, such as SO, efforts should be made to re-annotate the "legacy" sequence data using the standards.

### *Applicability to NCI*

Because it is under development, and to our knowledge no major genome database has explicitly indicated incorporation of SO at this time, it is not recommended for deployment at this time.

### *NCI role*

This standard is under development and not in use, so NCI should monitor its development.

# 3. HEALTH-RELATED TRANSACTION STANDARDS

## 3.1 Clinical Transaction Standards

This section addresses health-related transaction standards that specify content, definition, and format of health-related data exchanges.

### 3.1.1 American National Standards Institute (ANSI) Accredited Standards Committee X12 (X12)

<http://www.x12.org/x12org/subcommittees>

### *Sponsor*

Accredited Standards Committee X12 (ASC X12) is a standards organization that is accredited (hence the name) by the American National Standards Institute (ANSI). The organization was accredited in 1979 to develop uniform standards for inter-industry electronic exchange of business transactions-Electronic Data Interchange (EDI). X12 is organized into several subcommittees, each of which develops specifications within a given business context. The X12N (Insurance) is the relevant committee for messaging within the healthcare community.

### *Description*

The X12 organization has defined a common structure and encoding for all its specifications, and it maintains a common data dictionary that supports those specifications. At the highest level X12 is an EDI standard. The conceptual basis for X12 standards is replacing paper documents with electronic transmissions for supporting the flow of information between businesses and other organizations.

The X12N subcommittee has declared the following purpose and scope:

- "Develops and maintains X12 EDI and XML standards, standards interpretations and guidelines as they relate to all aspects of insurance and insurance-related business processes.
- Includes development and maintenance activities relating to property, casualty, health care, life, annuity, reinsurance, pensions, and reporting to regulatory agencies. Insurance Subcommittee initiatives also include all products and services, such as government health care programs like Medicare.
- Serves as a liaison with complementary insurance standards bodies, such as HL7, to coordinate standards development activities."[19]

[19] < http://www.x12.org/x12org/subcommittees >

### *Structural Elements of an X12 Standard*

X12 is a messaging standard that bears many similarities to Health Level 7 (HL7). In fact, it might better be said that HL7 bears similarity to X12 since the X12 standard was developed earlier and, given that its membership extends outside of the healthcare community, is more widely used. An X12 message is defined by reference to a *transaction set*, which is a collection of segments. Each *segment* is a collection of *data elements,* which are defined within a data element dictionary that specifies the format for the element and the allowable *code values* for coded elements. These key elements of the standard are discussed below:

- **Transaction set**: The transaction set defines the collection of segments that make up a valid segment. The transaction set is a nested collection of loops—each of which contains a list of segments and/or enclosed loops. Each transaction set supports a particular functional requirement, e.g., Patient Claim (837). Conceptually, the transaction set represents and replaces a paper document.
- **Segment**: The segment defines a collection of data elements to be used in different contexts. An X12 segment tends to be very fine grained so that it can be easily reused. For example, here is a list of segments used in the patient eligibility transaction set to indicate items for which eligibility is declared: Eligibility or Benefit Information EB, Health Care Services Delivery HSD, Reference Information REF, Date or Time or Period DTP, Request Validation AAA, Vehicle Information VEH, Product/Item Description PID, Property Description - Real PDR, Property Description - Personal PDP, Item Identification LIN, Equipment Characteristics EM. It is important to note that each of these segments (with the probable exception of the HSD) is used over and over in transaction sets whose usage spans from freight delivery, to banking, to healthcare.
- **Data Element**: The data element is an atomic item of information. The data element dictionary notes that, "The data element is the smallest named unit of information in the standard… For each data element, the dictionary specifies the name, definition, type, minimum length, and maximum length."[20] As with segments, data elements are designed in a generic fashion so they can be reused within segments that are applied across a wide variety of contexts.
- **Code Value**: All coded elements—those with type = ID—have their valid values listed within the data dictionary, or have a reference to the external source for the codes. Given that data elements are intended for reuse across transaction sets, most code sets include values that are used across the entire gamut of X12-supported industries.

### *Functional Scope of X12N*

In the healthcare arena, the transaction sets mandated by the Health Insurance Portability and Accountability Act of 1996 (HIPAA) are particularly significant. The following transaction sets (message specifications) are included within the scope of the HIPAA legislation:

- Payroll Deducted and Other Group Premium Payment for Insurance Products (820)[21]: Companies and government agencies that offer employees group life, health, and disability insurance can use a subset of the 820 to provide remittance detail associated with the

[20] Data Element Dictionary, Page *vii*; dpANS X12.3 – February 6, 2003

[21] This is the transaction set ID assigned by X12.

premium payments. The premium being remitted can be associated with health care, individual life, disability, and/or property and casualty contracts.[22]

- Health Care Claim Payment/Advice (835): The 835 is intended to meet the particular needs of the health care industry for the payment of claims and transfer of remittance information. The 835 can be used to make a payment, send an Explanation of Benefits (EOB) remittance advice, or make a payment and send an EOB remittance advice from a health care payer to a health care provider,
- Health Care Eligibility Benefit Inquiry and Response (270/271): The Health Care Coverage, Eligibility, and Benefit transactions are designed so that inquiry submitters (information receivers) can determine (a) whether an information source organization (e.g., payer, employer, HMO) has a particular subscriber or dependent on file, and (b) the health care eligibility and/or benefit information about that subscriber and/or dependent(s).
- Health Care Services Review – Request for Review and Response (278): The 278 has the flexibility to accommodate the exchange of information between providers and review entities.
- Benefit Enrollment and Maintenance (834): The 834 is used to transfer enrollment information from the sponsor of the insurance coverage, benefits, or policy to a payer.
- Health Care Claim (837): The 837 is used in developing and executing the electronic transfer of health encounter and health claim data. Note: X12 has provided separate implementation guides for institutional, dental, and professional claims.

X12N – The insurance subcommittee has developed several other transaction sets, listed below, that are relevant within the health care arena:

- Animal Toxicological Data (249).
- Health Care Benefit Coordination Verification (269).
- Health Care Provider Information (274).
- Patient Information (275). This transaction set is noteworthy because it provides for the encapsulation of an HL7 message as a mechanism for encoding detailed clinical information.
- Health Care Claim Status Request/Information (276/277).
- Medical Event Report (500).

### *Usage*

Currently X12 standards are widely used within the healthcare industry for reimbursement and insurance related messaging. This role of X12 messaging was initiated by the interest of payers in having consistency in receiving claims information from healthcare providers. The use of X12 has been greatly enhanced by its central role within the Health Insurance Portability and Accountability Act of 1996 (HIPAA). Through the HIPAA legislation, congress intended to encourage using industry standards to reduce the burden of administrative costs on the healthcare system.

### *Applicability to NCI*

As stated above, X12 offers a set of transaction definitions, along with segment and data definitions that support those transaction sets. However, it is important to recognize that the vocabularies and data definitions provided by X12 do not stand on their own—they are useful in the context of the applicable transaction set. Another way to explain this point is to note that,

[22] The descriptions of the transaction sets are taken from the Business Purpose section of the corresponding HIPAA Implementation Guides.

within X12 standards, the data elements (and the segments that are the first order of grouping) are designed to be as generic as possible. The semantics of an element or segment only becomes clear in the context of its use within a transaction set. Consequently, there is little point in storing these definitions, unless they will be used to support X12-based messaging.

Given its primary interest in clinical trials and health care research, standardizing communications between health care payers and health care providers would not seem to be an NCI area of interest. So, there will be negligible value to importing this data.

***Curation***

The data is available in Portable Document Format (PDF), paper, and database form. Therefore it would be easy to download. Note that X12 produces a new set of specifications every six months. However, because successive releases vary little, there is no formal requirement for backwards compatibility. This state of affairs would require NCI to process periodic downloads to stay current.

Since there is a fee associated with access to the most current version, it is unlikely that NCI would be able to make the data available in the publicly accessible caDSR. NCI would only be able to register that data that is made available on the X12 web site.

***NCI Role***

Membership in ASC X112 is open to any individual, company, or organization that may be directly and materially affected by ASC X12 activities. There is an annual dues payment required for membership. The membership dues are based on the size of the organization, so they might be substantial. However, there is minimal benefit to NCI in joining ASC X12, unless there are areas of standardization in which NCI would like to participate.

### 3.1.2. American Society for Testing and Materials (ASTM) ASTM E1384-02a Standard Guide for Content and Structure of the Electronic Health Record (EHR)

<http://www.astm.org/>

***Sponsor***

ASTM E31.19 produces the standard, which is a working group with the American Society for Testing and Materials (ASTM) Committee E31. E31 is the ASTM committee that addresses health informatics. The committee website notes: “ASTM Committee E31 on Healthcare Informatics develops standards related to the architecture, content, storage, security, confidentiality, functionality, and communication of information used within healthcare and healthcare decision making, including patient-specific information and knowledge.”[23] The website also notes that committee E31 currently has 70 members and meets twice a year.

***Description***

The standard aims to:

[23] http://www.astm.org/commit/committee/E31.org

1. "identify the content and logical structure of an Electronic Health Record (EHR) consistent with currently acknowledged patient record content.
2. Explain the relationship of data coming from diverse sources … and other data in the Electronic Health Record as the primary repository for information from various sources.
3. Provide a common vocabulary for those developing, purchasing, and implementing EHR systems.
4. Provide sufficient content from which data extracts can be compiled to create unique setting 'views.'"[24]

The document consists of a general discussion of electronic health records, and of a data dictionary that includes a list of the data elements to be included within an electronic health record system. The discussion includes a high-level object model depicting some of the key elements of the patient record, and a fairly useful table categorizing different types of data that are included in a typical record.

The data dictionary includes "all elements proposed for the record, including primary and longitudinal records from multiple sources. They are listed as a visual summary only to aid perception of the complete pattern of the record. Each data element is also listed and further detailed with attributes in Annex A1, that is part of this guide. The reader is again reminded that, though each data element characterization is a part of this guide, it is not required if it is not to be used, but, if used it must have the same meaning and representation as given in Annex A1."[25] For each entry in Annex A1, the standard supplies an attribute name, a formalized name that provides an indication of type of data that is represented, and a brief description of the element. The listing of data elements is suggestive rather than comprehensive—perhaps a comprehensive list would not be feasible.

The standard also includes a set of code value tables that provides sets of values for many of the coded elements within the data dictionary. These code tables should be reviewed for possible inclusion in the NCI repositories. However, several of the tables, e.g., race, occupation, are drawn from other sources that are covered in this review. In such cases, it will be better to draw the code set from its original source.

### *Usage*

The EHR standard provides general, qualitative guidance to issues that should be considered in constructing an electronic health record system, and offers suggestions regarding the data to be supported by such a system. The standard does not offer guidance as to what would constitute conformance with its recommendations, and much of the data recommended is data that systems would need to support without reference to the standard. As a result, it is very hard to gauge the extent to which the standard is used within the U.S. healthcare environment.

### *Applicability to NCI*

The EHR standard is not directly relevant to the day-to-day activities of the NCI. However, the standard's discussion of typical and important data for various aspects of patient management and care is relevant, because the included data dictionary can be used as a checklist of data that is relevant in clinical trials and other patient care-related situations.

[24] Section 1, Page 1; ASTM E1384-02a; ASTM International, 2003
[25] E1384-02a opp. Cit. Page 23-24.

### *Curation*

If it becomes necessary to register the standard within the NCI repositories, this will be a relatively labor-intensive process since the standard is only provided as a printed document.

### *NCI Role*

ASTM is an accredited standards organization within the ANSI context. There are several categories of membership available that allow organizations or individuals to attend meetings and to participate in developing standards.

Collecting this data occurs periodically and uses various instruments. Data is collected though direct patient interviews, through telephone interviews, through extraction from medical and provider records, and through direct examination at a mobile examination site. NCHS provides documentation of the data that is collected by providing documentation on the questions that are asked, the layout of the data files made available to researchers, and through specifying the criteria used to constrain the data collected for a particular item. In some cases, data items are constrained through specification that responses must be drawn from a particular code set—in these cases the valid codes are defined directly within the survey and data output documentation. There does not seem to be a central representation of the attributes collected across the body of surveys.

## 3.2. Genomics Transaction Standards

## 3.2.1 Biomolecular Sequence Analysis (BSA) Specification

<http://www.omg.org/technology/documents/formal/biomolecular_sequence.htm>

### *Sponsor*

The BSA specification was developed and submitted by the following industry and academic institutions to the Object Management Group (OMG): Concept Five Technologies, Inc.; EMBL-EBI (European Bioinformatics Institute); Genome Informatics Corporation; Millennium Pharmaceuticals, Inc.; Neomorphic Software, Inc.; NetGenics, Inc.; Oxford Molecular Group; Sanger Centre.

### *Description*

The BSA specification contains two main modules, BioObjects and Analysis. The BioObjects module defines the elements that can be used to represent properties of biological sequences, such as plus and minus strands for nucleotide sequence, genetic code used, start and end of a sequence region that a particular feature or annotation is associated with, and sequence alignments. The Analysis module specifies the elements that can be used to represent sequence analysis events, such as type of analysis (e.g. BLAST, pattern finding, etc), analysis state, input/output format, job control, etc.

There are overlapping data models such as those developed by NCBI, and few databases claim to be using BSA.

***License***

As an object specification, it is open and downloadable to the public from the OMG. Implementations of the BSA by different software tool developers may adopt different licenses. An open source project to implement the BSA (OpenBSA) is available from EBI (http://industry.ebi.ac.uk/openBSA/), but has not been updated since 09/2000.

***Usage***

The BSA specification can be used as a guide to design a data model to represent biological sequences and analysis. The BSA also defines an interface format so that different tools can inter-operate and exchange data, as long as these tools are compliant with the BSA specification.

The difference between the BSA specification and the Sequence Ontology (SO) is that the latter provides the terms or content that can be used to provide values for the elements that are defined by the BSA. However, we are not aware of any public database that implemented the BSA standard.

***Applicability to NCI***

As this is not in use, it is not clear whether this will be applicable to NCI.

***NCI role***

NCI can monitor its use as a potential source for a data model for genomic and proteomic sequences for an informatics system.

### 3.2.2 Genomic Maps Specification <http://www.omg.org/technology/documents/formal/genomic_maps.htm>

***Sponsor***

This standard was submitted to the OMG by EMBL-EBI (European Bioinformatics Institute); Millennium Pharmaceuticals, Inc.; NetGenics, Inc.

***Description***

The Genomic Maps standard defines a data model to represent genomic maps and their contents. The specification contains three modules, “ControlledVocabularies”, “GenomicMaps”, and “LQSlink” and the compliance points, which define 4 levels and requirements of implementation for each level.

The “ControlledVocabularies” module defines a structure, such as VocabularyString, VocabularyEntry, etc., to bring in domain specific genomic vocabularies, e.g. taxonomy, the Sequence Ontology (SO). The “LQSlink” module provides a mechanism to connect the vocabularies to the Lexicon Query Service (LQS), an OMG specification for accessing the content of medical terminology systems. The “GenomicMaps” module defines a structure to represent the elements that can be mapped to the genome, their start/end points, and length. Overlapping data models can be found in NCBI, and other genome databases, however, few claim to be compliant with the Genomic Maps Specification.

***License***

The Genomic Maps Specification is an open standard and can be downloaded from the OMG website.

***Usage***

The Genome Maps Specification can be used for designing genomic mapping database schema, that provide a flexible design for utilizing already developed standards such as controlled vocabularies and ontologies. To our current knowledge, none of the major genomic databases has implemented the Genomic Maps standard for production database operation.

***Applicability to NCI***

NCI already has data model to accommodate genomic mapping information. However, compliance of NCI data model to Genome Maps Specification would facilitate exchange and sharing of genome mapping information between different organizations with minimal efforts. However, as it is not in use at this time, NCI deployment is not advised.

***NCI role***

Monitor the usage in the community.

### 3.2.3 Standard for Exchange of Nonclinical Data (SEND)

<http://www.pharmquest.com/send_consortium/overview.htm>

***Sponsor***

The SEND consortium is hosted by PharmQuest with representatives from the pharmaceutical industry, contract research organizations (CROs), software vendors and regulatory agencies (FDA's Center for Drug Evaluation & Research - CDER, Center for Food Safety and Nutrition - CFSAN and Center for Veterinary Medicines - CVM).

***Description***

SEND is a data standard for the industry to submit non-clinical data to the FDA as a part of submission for a new drug, biological, veterinary medicine or food product. The focus of the SEND Consortium has been on data collected from animal Toxicology studies. SEND was developed using similar rules and principles as the Clinical Data Interchange Standard Consortium (CDISC).

SEND contains a broad range of domains that cover subject areas of exposure, animal characteristics, animal disposition, body weights, clinical pathology, clinical signs, drug/metabolite levels, food consumption, fetal data, female fertility, group characteristics, group observations, macroscopic findings, male fertility, microscopic findings, ophthalmoscopic findings, organ weights, rodent micronucleus, study summary, tumor analysis, water consumption, and study timing. A dataset is a flat file with one or more rows and columns, containing all the observations for each domain of a study. Each row of a dataset represents a single observation while each column represents a variable. A variable is defined with seven metadata attributes: variable name, descriptive variable label, data type, controlled terminology, origin, role (or how the variable is used), and comments. SEND also provides standard dataset attributes (or SEND models) for the domains that provide standard variable names and controlled terminology. For example, to describe the type of a body weight measurement, one should use Variable_Name::BWRESTYP, Variable_Label::Body Weight Result Type, Type::Char, Controlled_Terms::Baseline (or Intermediate, or Final, or Terminal).

### *License*

SEND is an open standard that can be downloaded from the CDISC website.

### *Usage*

The focus of the SEND Consortium has been on data collected from animal Toxicology studies. SEND is intended to facilitate transfer of non-clinical data from sponsor to the FDA and subsequent loading into the FDA repository.  The extent of its usage is unknown.

### *Applicability to NCI*

Since SEND is a standard adopted by FDA, it should be considered by NCI in the future

### *NCI role*

The NCI currently has no role in this standard, and is not tracking it.  It may be useful to keep track of its status and usage in the community.

# APPENDIX D - Standards Review And Approval Bodies

## 1. CENTERS FOR DISEASE CONTROL, PUBLIC HEALTH INFORMATION NETWORK (PHIN) <http://www.cdc.gov/phin/>

### *Sponsor*

The PHIN is sponsored by the Centers of Disease Control and Prevention (CDC) in order to provide a "crosscutting and unifying framework to monitor clinical data streams for early detection of public health issues and emergencies."[26] PHIN is an expanded and updated successor to the National Electronic Disease Surveillance System (NEDSS), which had a primary focus on standardization and modernization of a large and disparate collection of independent disease surveillance systems.

### *Description*

The CDC Web site, provides the following description of PHIN. "Through defined data and vocabulary standards and strong collaborative relationships, the Public Health Information Network will enable consistent exchange of response, health, and disease tracking data between public health partners. Ensuring the security of this information is also critical as is the ability of the network to work reliably in times of national crisis. PHIN is composed of five key components: detection and monitoring, data analysis, knowledge management, alerting and response:

- **Detection and Monitoring**
  Focus: Disease and threat surveillance, national health status indicators.
- **Analysis**
  Focus: Facilitates real-time evaluation of live data feeds, turning data into information for people at all levels of public health.
- **Information Resources and Knowledge Management**
  Focus: Providing intuitive access to reference materials, integrated distance learning content and decision support.
- **Alerting and Communications**
  Focus: Enabling emergency alerting, routine professional discussions and collaborative activities.
- **Response**
  Focus: Management support of recommendations, prophylaxis, vaccination, etc."[27]

It is important to recognize that CDC has taken on a very large and challenging systems development project in announcing, promulgating, and working to implement systems according to PHIN standards. The scope of PHIN extends across 80-120 disease and condition surveillance program areas, and covers some 55 jurisdictions reporting directly to CDC (This includes U.S. states, some large metropolitan entities, and several U.S. territories). The impact of PHIN, and the changes envisioned by PHIN developers, extends beyond the impressively large number of systems implied by the numbers above to include local public health entities, as well as the laboratories and providers that provide data. Given this massive scale, PHIN must be evaluated as a work in progress, many of whose parts are either in conception, still in development, or in the process of rolling out.

[26] CDC web site, http://www.cdc.gov/phin/
[27] Ibid

Efforts under the PHIN umbrella include: a) development of vocabulary and messaging standards, b) development of data transmission and management standards, c) creation of standards for data display and entry, d) and the implementation of application and databases to implement these standards. The first of these headings is important within this context.

### *Messaging Standards*

PHIN messaging standards are divided into two categories: a) transactions within the public health sector, and b) transactions that convey information from healthcare providers and laboratories to public health departments. This distinction is significant because of the standards focus that CDC has adopted for the PHIN. CDC has chosen to standardize on HL7 for data exchange in the clinical and public health arenas.[28]

The following is a list of current PHIN sponsored implementation guides that use the HL7 Version 2 standards:

- Implementation Guide for Transmission of Laboratory-Based Reporting of Public Health Information: Electronic messaging of most laboratory-reportable findings associated with notifiable disease conditions from laboratories to public health agencies in accordance with the HL7 standard.
- Implementation Guide for Transmission of Microbiology Result Reporting of Public Health. The guide, which complements the Electronic Laboratory Reporting Implementation Guide, is primarily intended for rapid and early reporting of microbiology results, such as initial growth of an organism in a blood or sputum culture prior to identification of a specific organism.
- Implementation Guide for Transmission of Laboratory, Pharmacy and Supply Orders as Public Health HL7 clinical reporting to allow health care providers to transmit information about laboratory, pharmacy, and supply orders issued for patient care to public health agencies.
- Implementation Guide for Transmission of Patient Chief Complaint as Public Health HL7 patient management reporting to allow health care providers to transmit information about chief complaints for new inpatient, outpatient, and emergency encounters to public health agencies.
- Laboratory Result Message for Bio-terrorism Response Health Level 7 (HL7) Version 2.4 ORU (Unsolicited Observation Message) to support reporting laboratory results in the context of Bio-terrorism response messaging.

The following list includes the implementation guides that have been created using the Version 3 standards. Each document provides a reference for implementing Notification Reporting under the PHIN architecture for a particular disease entity, or group of related disease entities.

- *Neisseria meningitides* Case Notification
- *Haemophilus influenzae* Case Notification
- Group B *Streptococcus* Case Notification
- Group A *Streptococcus* Case Notification
- *Streptococcus pneumoniae* Case Notification

[28] The decision to standardize on HL7 predates the advent of PHIN. HL7 standards have been created for the messaging that supports Immunization Registries, and there has been a long standing Implementation Guide to support laboratory reporting to public health departments.

- Other Meningitis Case Notification
- Hepatitis A Acute Case Notification
- Hepatitis B Acute Case Notification
- Hepatitis C Acute Case Notification
- Perinatal Hepatitis B Case Notification
- Chronic/Resolved Hepatitis C Case Notification
- Hepatitis Non-ABC, Chronic Hepatitis B Case Notification
- Measles Case Notification
- Rubella Case Reporting
- Pertussis Case Notification
- Generic Case Notification
- Summary Case Notification

This document provides a reference for implementing Notification Reporting under the PHIN architecture for a particular disease entity, or group of related disease entities. This particular implementation guide is designed to support summary case reporting across all disease entities. The summary guide provides information about the number of cases of a particular disease or condition that have been reported during a particular time period.

***Vocabulary Standards***

PHIN developers are making use of specified vocabularies in two different but related contexts. The contexts are a) the development of the NEDSS Base System and the creation of PHIN sponsored message specifications. Where possible the value sets that are specified are based on those published by HL7; this is a consequence of PHIN's decision to use HL7 as its base clinical standard. However, in those situations in which HL7 has not developed a value set, or in which public health has particular requirements, PHIN has developed and plans to maintain its own vocabulary items. It is intended that these, more particular, domains and value sets be submitted to HL7 so they can be included within the HL7 vocabularies.

The list of PHIN supported vocabulary domains is quite long. At this point time, they, for the most part, include the value sets needed to support infectious disease surveillance and reporting. Over time, the scope will expand to go beyond infectious disease to cover areas such as environmental health.

***Usage***

The primary user of the PHIN standards is the CDC itself, in its role as a developer of systems for federal and state use. CDC is developing (through a collection of contractors) the NEDSS Base System to offer a PHIN conformant software package to states that do not intend to develop their own surveillance and case management systems. CDC is also developing systems to be used by CDC departments and program areas for a) their own surveillance, reporting, and disease management activities, and b) to support bio-terrorist detection and response.

The secondary users include states that develop their own systems for surveillance and reporting to CDC. The states are expected to follow PHIN standards for managing surveillance data, and they are expected to follow PHIN standards for data exchange while reporting to CDC. PHIN standards will also affect other parties who provide surveillance and monitoring data to CDC and to public health departments. Such parties include public health labs and healthcare providers. Within this context, HL7 Version 3 specifications are being created for messaging between state health departments and CDC, while Version 2 specifications are recommended for messaging from healthcare providers and labs to health departments.

### *Applicability to NCI*

PHIN standards are, and will continue to be, based on those developed by HL7. These standards will over time replace the localized standards that are currently used in public health reporting and data management. In particular, it is expected that PHIN developers will create structures to manage and distribute PHIN vocabulary value sets, and that the scope of these value sets will grow to include all the areas where CDC manages data. At this time, the focus is on infectious disease reporting, not cancer. As a result, NCI should consider PHIN as a secondary source for vocabulary items (This recommendation will be most useful in a context where HL7 is seen as a primary source for most non-specialized vocabularies.) To the extent that NCI is interested in the vocabularies used in the surveillance process, PHIN will become a primary source.

Since PHIN will continue to place great reliance on PHIN standards, PHIN developers will look first to HL7 as a source of vocabulary items. Secondarily, PHIN developers plan to funnel the vocabularies that they produce into HL7 to provide wider use and acceptability for those domains and value sets. In the future a web-enabled vocabulary registry may support PHIN and might call for some coordination with EVS.

### *Potential for Use*

Those PHIN vocabularies that do not appear in HL7 could be registered within the caDSR and/or incorporated into EVS. Care would need to be taken that the same items are not registered multiple times when they appear both in PHIN and in HL7. On the other hand, there is little point in registering the PHIN messages and implementation guides since these structures are only relevant in the context of the specific information flows they are designed to implement. To the extent that the data structures underlying these specifications are relevant, those data structures can be captured through registration of the HL7 Reference Information Model.

### *NCI Role*

PHIN is a federal government initiative managed by CDC, not a standards organization. As such, it does not have defined structures for allowing input from parties outside of the direct PHIN orbit. Furthermore, there does not seem to be much of a role for NCI input unless the kinds of data flow that PHIN manages converges with those supported by NCI. Currently this is not the case, but functional overlaps could develop, especially as PHIN tackles surveillance and reporting of chronic diseases such as cancer. If there is an interest in having NCI involvement in PHIN, the first step would be to create institutional contacts between parties within NCI and the Information Resource Management Organization (IRMO) within CDC.

## 2. CONSOLIDATED HEALTH INFORMATICS INITIATIVE (CHI)

<http://www.whitehouse.gov/omb/egov/gtob/health_informatics.htm>

The Consolidated Health Informatics (CHI) initiative aims to "establish a portfolio of existing clinical vocabularies and messaging standards enabling federal agencies to build interoperable federal health data systems." CHI standards will work in conjunction with the Health Insurance Portability and Accountability Act (HIPAA) transaction records and code sets and HIPAA security and privacy provisions. About 20 department/agencies including HHS, Department of Veterans Affairs (VA), DoD, Social Security Administration (SSA), General Services Administration (GSA), and the National Institute of Standards and Technology (NIST) are active in the CHI governance process. Through this process, all federal agencies will incorporate the adopted standards into their individual agency health data enterprise architecture used to build all new systems or modify existing ones. As of May 2004, the following standards had been approved for adoption:

- Health Level 7 (HL7) messaging standards to ensure that each federal agency can share information that will improve coordinated care for patients such as entries of orders, scheduling appointments and tests and better coordination of the admittance, discharge and transfer of patients.
- National Council on Prescription Drug Programs (NCPDP) standards for ordering drugs from retail pharmacies to standardize information between health care providers and the pharmacies. These standards already have been adopted under the Health Insurance Portability and Accountability Act (HIPAA) of 1996.
- The Institute of Electrical and Electronics Engineers 1073 (IEEE1073) series of standards that allows for health care providers to plug medical devices into information and computer systems so health care providers can monitor information from an intensive care unit (ICU) or through telehealth services on Indian reservations, and in other circumstances.
- Digital Imaging Communications in Medicine (DICOM) standards that enable images and associated diagnostic information to be retrieved and transferred from various manufacturers' devices as well as medical staff workstations.
- Laboratory Logical Observation Identifier Name Codes (LOINC) to standardize the electronic exchange of clinical laboratory results.
- Health Level 7® (HL7®) vocabulary standards for demographic information, units of measure, immunizations, and clinical encounters, and HL7®'s Clinical Document Architecture standard for text based reports. (Five standards)
- The College of American Pathologists Systematized Nomenclature of Medicine Clinical Terms® (SNOMED CT®) for laboratory result contents, non-laboratory interventions and procedures, anatomy, diagnosis and problems, and nursing. HHS is making SNOMED-CT® available for use in the U.S. at no charge to users. (Five standards)
- Laboratory Logical Observation Identifier Name Codes® (LOINC®) to standardize the electronic exchange of laboratory test orders and drug label section headers. (One standard.)
- The Health Insurance Portability and Accountability Act (HIPAA) transactions and code sets for electronic exchange of health related information to perform billing or administrative functions. These are the same standards now required under HIPAA for health plans, health care clearinghouses and those health care providers who engage in certain electronic transactions. (One standard.)
- A set of federal terminologies related to medications, including the Food and Drug Administration's names and codes for ingredients, manufactured dosage forms, drug products and medication packages, the National Library of Medicine's RxNORM for describing

clinical drugs, and the Veterans Administration's National Drug File Reference Terminology (NDF-RT) for specific drug classifications. (One standard.)

- The Human Gene Nomenclature (HUGN) for exchanging information regarding the role of genes in biomedical research in the federal health sector. (One standard.)
- The Environmental Protection Agency's Substance Registry System for non- medicinal chemicals of importance to health care. (One standard.)
- The National Cancer Institute's Anatomy component of the NCI Thesaurus, which extends present anatomy terminologies into subcelluar structures that are required for research and internationally based clinical trials, and which is primarily recommended for that purpose. Additionally, the remaining Anatomy terminology may serve as an alternate for SNOMED CT. It is recommended that the two terminologies be related through mapping. (One standard.)

CHI also is reviewing recommendations in the following domains: Disability, History and Physical, Medical Devices and Supplies, Multimedia, and Population Health.

## 3. FOOD AND DRUG ADMINISTRATION (FDA) CENTER FOR DRUG EVALUATION AND RESEARCH (CDER)

<http://www.fda.gov/cder/dsm/>

The U.S. Food and Drug Administration, Center for Drug Evaluation and Research (CDER), has compiled the CDER Data Standards Manual (DSM), a compilation of standardized nomenclature monographs that have been reviewed and approved by the CDER Nomenclature Standards Committee (NSC). DSM monographs may have been derived either wholly or in part from other nomenclature standards settings bodies, as well, such as the International Conference on Harmonization (ICH), the United States Pharmacopeia (USP), the United States Adopted Names Council (USAN), the American Hospital Formulary Service (AHFS), the Chemical Abstracts Service (CAS), the National Institutes of Standards and Technology (NIST), the International Organization for Standardization (ISO), the American Society for Testing and Materials (ASTM), US Census Bureau, US Postal Service, and the Central Intelligence Agency.

FDA CDER has selected standards for commonly used data such as age, race, date, marital status, and telephone number, as well as the more specific topics of drug nomenclature, including drug classification, dosage form, route of administration, ingredient name, and the like.

This set of vocabularies has served as a source for some of the Common Data Elements and their permissible value sets that have been registered in the caDSR. These data sets are also being worked on by other standards bodies. NCI will evaluate them individually for possible use.

## 4. NATIONAL COUNCIL ON VITAL AND HEALTH STATISTICS (NCVHS)

<www.ncvhs.hhs.gov/>

### *Sponsor*

The National Council on Vital and Health Statistics is sponsored by HHS. "The National Committee on Vital and Health Statistics is the Department's statutory public advisory body on health data, statistics and national health information policy. This Committee shall serve as a national forum on health data and information systems. It is intended to be a forum for the collaboration of interested parties to accelerate the evolution of public and private health information systems toward more uniform, shared data standards, operating within a framework protecting privacy and security. The committee shall encourage the evolution of a shared, public/private national health information infrastructure that will promote the availability of valid, credible, timely and comparable health data. With sensitivity to policy considerations and priorities, the committee will provide scientific-technical advice and guidance regarding the design and operation of health statistics and information systems and services and on coordination of health data requirements. The Committee also shall assist and advise the department in implementing the Administrative Simplification provisions of the Health Insurance Portability and Accountability Act (HIPPA), and shall inform decision making about data policy by HHS, states, local governments, and the private sector."[29]

### *Description*

NCVHS does not create or maintain standards. However, it does evaluate standards and make recommendations regarding those standards that should be encouraged for use within the context of the Consolidated Healthcare Informatics Initiative (CHI). The NCVHS Web site notes that:

"The Committee provides advice and assistance to the Department and serves as a forum for interaction with interested private sector groups on a variety of key health data issues.

The Committee is composed of 18 individuals from the private sector who have distinguished themselves in the fields of health statistics, electronic interchange of health care information, privacy and security of electronic information, population-based public health, purchasing or financing health care services, integrated computerized health information systems, health services research, consumer interests in health information, health data standards, epidemiology, and the provision of health services. The Secretary of HHS appoints sixteen members for terms of four years each; with about four new members being appointed each year. Two additional members are selected by Congress."[30]

The NCVHS recommendations to the Secretary of HHS are quite relevant in the context of the NCI review of external standards. In its letter to the secretary regarding vocabulary standards the committee stated:

"NCVHS recommends that the federal government recognize a 'core set' of Patient Medical Record Information (PMRI) Terminologies as a national standard. This core set should comprise the minimal set of terminologies that (1) are required to adequately cover the domain of patient medical record information and (2) meet essential technical criteria to serve as *reference terminologies*. Furthermore, the NCVHS recommends that these terminologies be organized into a coherent, internally consistent, minimally redundant, cross-referenced core terminology set. Also, the NCVHS recommends that you designate the National Library of Medicine (NLM) as a central coordinating body to manage this terminology resource and coordinate its ongoing maintenance and distribution. The initial terminologies recommended for the core set of PMRI terminology standards are:

- SNOMED CT (as licensed by the National Library of Medicine)
- Logical Observation Identifiers Names and Codes (laboratory subset)
- Federal Drug Terminologies:
  - RxNorm

[29] Charter, National Council on Vital and Health Statistics. < http://www.mcvhs.hhs.gov/01charter.htm >.

[30]< http://www.ncvhs.hhs.gov/ >

- The representations of the mechanism of action and physiologic effect of drugs from NDF-RT
- Ingredient name, manufactured dosage form and package type from the FDA."[31]

***Usage***

Recommendations from NCVHS carry wide ranging authority based on the highly respected membership of the committee, and on its statutory and significant role within HHS. On the other hand, the different implications for healthcare providers of implementing these standards have not been fully worked out, and it is clear that migration to thoroughgoing use of vocabulary and messaging standards implies sweeping and expensive changes to current processes and information systems. It is reasonable to expect that parties within the healthcare system will increase their use of the NCVHS recommended standards. They will do so in the context of other requirements that will often be more immediately pressing than moving to support industry standards.

***Applicability to NCI***

NCVHS recommendations are highly applicable to NCI. The selection of a constrained set of standards can be used as a model for NCI because the NCVHS recommendations are based on the appreciation of choosing an appropriate and limited set of standards for support. The most recent NCVHS recommendation for the CHI Initiative should be used as a reference to define potentially applicable standards <http://www.ncvhs.hhs.gov/040129lt.pdf >.

***NCI Role***

In its role as a public advisory body, the NCVHS places great importance on holding public meetings and on receiving input from informed parties. At the same time, the committee membership includes liaison representation from the Centers for Medicare and Medicaid Services, the Agency for Healthcare Research and Quality, Centers for Disease Control and Prevention, the National Institutes of Health, and the National Center for Health Statistics. The NCI has both provided formal testimony to the committee and worked within the context of the liaison representation to provide input on recommended standards to the NCVHS.

[31] Recommendation for PMRI Terminology Standards, p.3.