Biomarkers Knowledge System

Meeting Report
September 8, 2000

Office of Science Policy
Office of the Director, NIH
Bethesda, MD

Meeting Summary Report

Meeting Summary Report
Biomarkers Knowledge System
National Institutes of Health
Bethesda, MD
September 8, 2000

Introductory Remarks/Goals and Objectives:

The meeting commenced with introductory remarks from Lana Skirboll, Director of the Office of Science Policy at the NIH. Dr. Skirboll noted that recent developments in fields such as proteomics and the "genetics revolution" have produced volumes of data about biomarkers that must be standardized and properly recorded in order to maximize their usefulness. Meeting participants were then charged with the task of examining possible approaches to the handling and dissemination of data between the various members of the scientific, technical and clinical communities. A possible long term goal of this meeting is to focus on pathways that facilitate information transfer throughout the research community in order to inform health care practices.

The following definitions were provided for the meeting discussions:

Biomarker (biological marker): A characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to therapeutic interventions.

Knowledge System: An online (electronic) source and linkage of databases, navigation and analysis tools, and descriptive information about data having attributes such as a structure for conducting queries and analyses, support for browsing and retrieving information, a cognitive structure for interpreting data, and the capability for online conversations or exchange of data and information.

Several current difficulties associated with biomarker information were noted, including:

- A lack of systematic organization—data are published in too many places
- Data are not presented uniformly across datasets
- An increase in both the complexity of data and the number of variables
- Incomplete datasets
- Unpublished data
- An inability to associate with population data
- Lack of interoperability issues among datasets
- Non-uniform taxonomy and definitions

He then stressed that the goal of the meeting was to identify critical issues and discuss needs that will facilitate the exchange, archiving, and retrieval of knowledge about biomarkers to benefit the research community and ultimately inform health care practices. Specific meeting objectives included:

- ♦ Addressing technical and management issues associated with the integration of research databases and information in a biomarkers knowledge system
- ♦ Developing data architecture to support the creation of biomarker metadata
- ♦ Discussing approaches to creating common data elements, data dictionaries, and standards
- ♦ Coordinating biomarker knowledge system development with other information and standards organizations

Database issues that were not addressed specifically in this meeting included issues of data accuracy, database security and encryption, privacy and confidentiality concerns, intellectual property, shared versus retained data, and broader issues of public and private genomics and proteomics databases. This meeting therefore consisted of five major areas of discussion: the knowledge environment, research networks, data architecture, standards, and common data elements.

Session #1: Knowledge Environments—Lessons Learned, New Directions:

Speakers: **Michael Stout**, Oxford University Press; **Monica Bradford**, American Association for the Advancement of Science
Discussants: **Rochelle Long**, National Institute of General Medical Sciences; **Stephen Maurer**, University of California-Berkeley

Mr. Stout discussed the development of the knowledge environment, Cancer Spectrum, a cooperative research and development agreement (CRADA) between the Oxford University Press (OUP) and the National Cancer Institute (NCI). The CRADA was established in 1996 to run for five years for the purpose of publishing the *Journal of the National Cancer Institute* (*JNCI*) http://jnci.oupjournals.org/ in January 1997. Staffing for the duration of the CRADA is represented by both the government and the OUP. Under this agreement, the OUP has the following duties: to maintain high standards for the *JNCI*, to develop the *JNCI* in electronic form, and to develop a cancer knowledge environment around the *JNCI*. In 1999, the full *JNCI* on-line service was launched featuring full-text SGML articles (for searching and linking), PubMed links, links to full-text articles in other OUP and Highwire journals, and full-text PDF articles (for printing). Mr. Stout stressed the importance of market research for identifying the target market, designing the most useful user interface, and deciding on the content for Cancer Spectrum.

The Cancer Spectrum knowledge environment (KE) features a wide-ranging content including extended news coverage, NCI resources (SEER statistics, PDQ, Cancerlit, and a universal database), PubMed, OUP resources (journals, textbooks), International Agency for Research on Cancer resources (cancer registry statistics, WHO monographs summaries, IARC books, and Cancer Incidence in Five Continents and Globocan), and CanQuest. Users currently must register to access the Cancer Spectrum demonstrator system at http://cancerspectrum.oupjournals.org The Cancer Spectrum knowledge environment features include filtering by region, currency, topic, content type, and level, and concept searching by theme extraction, content categorization, content taxonomy, and latching rules. Cancer taxonomies are needed for browsing by topic, organizing statistics, and organizing useful links. Strategies to organize the available information include cancer taxonomy as the intersection of disease, factors, and onset taxonomies. Current issues in content integration include extending the technology to combine multiple taxonomies, allowing for permutations in terms within a theme, latching themes together based on context over several levels in the taxonomy, consolidating concepts, and presenting results.

The overall goal of the Cancer Spectrum KE is to convert data into knowledge by developing simple organizational structures to classify extracted themes and developing rules for combining simple knowledge structures to obtain new facts. This will then facilitate the transfer of information rather than data. Possible roles of the OUP in the biomarkers project include an overlap between Cancer Spectrum and Protein Profile in addition to OUP's potential as an experienced KE developer, a Content partner, and as an interested observer.

Monica Bradford focused on the American Association for the Advancement of Science's Signal Transduction Knowledge Environment (STKE), an online resource designed to take advantage of new technologies that will facilitate researchers' access to vital information and enhance the organization of that information to promote the transition from isolated facts to an integrated knowledge base. The goals of the knowledge environment include creating new tools for information management, leveraging new on-line technologies to link materials, capturing information that answers questions such as “how to,” “what is,” “where is,” and “who is,” and enhancing community building. The STKE is based on the concept that much of the most exciting science occurs at the interface between disciplines. Therefore, the KE uses an Internet-based work environment to collect and organize data-based information from numerous fields of inquiry. The knowledge environment concept arose in response to existing information-laden databases such as PubMed as well as the often uneven quality of web-based information sites. The purpose of the STKE was to supply the needed pieces of information in order for researchers to communicate across disciplines.

The rapidly moving field of signal transduction was chosen as the focus in the initial 1996 STKE partnership between the AAAS and Highwire Press. The field serves as a good test model for a KE because it generates volumes of information, is an acronym-heavy discipline, and represents an area of expertise for *Science*. Questions poised by the creation of the STKE included:

- Quality control—How to best search contexts and link material?
- User behaviors—would a KE actually save the user time and resources?
- How to choose between comprehensive display and editorial selection?
- How to foster communication between scientists, publishers, and the informatics community?
- How to attract the experts in the field?
- What tools are necessary to provide to the expert contributor?
- How to educate the community into the value of these electronic contributions?

The site features a drag and drop graphing tool, links, a controlled vocabulary from which the user can make a selection, and a citation manager. The environment is interactive, and the user can add information. In order to populate the database, authorities in the field (as assessed through the literature) contribute canonical protein pathways from which specific pathways can be developed. In the instance of multiple pathways, a connections map site allows for interfacing of pathways, and the inter-pathway connection will hopefully facilitate slight differences in the authorities’ pathways. Specific user information for the STKE can be found on-line at http://www.stke.org/misc/intro.dtl.

Discussion: Dr. Rochelle Long:

Dr. Long compared the STKE with the Pharmacogenetics Research Knowledge Base (PharmGKB), a venture evolving not from a print source but rather from large, multi-disciplinary groups convening to create a knowledge base. The PharmGKB was created to organize disparate information in a central location while insuring promoted collaborations between experts. The PharmGKB is comprised of working groups from fields including cancer, asthma, depression, transporters and the cytochromes P450 and involves a peer-reviewed research mechanism by which research is funded through RFA submission. The history of research fuels current work, and groups meet through a steering committee. The PharmGKB also funds an attorney to investigate the ethical implications of proposed research. Issues facing the PharmGKB include standardization of vocabulary, links to standardized resources, formation of a database to drive the research, and the formulation of tools to promote research and get information into the public domain. The formation of the PharmGKB involved careful discussion of who will use the resource, what content to feature, how to link information with other resources, and the motivating factors responsible for initially drawing people to the database.

Discussion: Stephen Maurer

Mr. Maurer responded by noting that these tools are bibliographic and should carry a disclaimer stating such. He emphasized the importance of human contribution to the informatics process and the need for models such as the STKE to continue incorporating human judgment. By contrast, he noted that the PharmGKB uses an academic model that features community consensus through steering committees, and he forecasted two possible outcomes of such a system: 1) participants may build highly specific databases of limited applicability or 2) participants can construct one reasonable operating system such that outside investigators will be encouraged to participate. Since the ultimate goal is to extend the database to the larger research community, an arrangement such as the STKE serves as a good model. Although the authorities meet to set the standards for the STKE, outside persons may critique their judgment, and this scenario keeps the system from becoming closed or provincial. Dr. Maurer also noted that market-based organizations such as the OUP have historically done well when making information available to the public.

Dr. Maurer also commented on options for financing such a knowledge environment. He noted that the overall goal is to create a knowledge environment that is both self-supporting and based on a community model. If the government provides some of the seed money for such a project, then it should have a voice in the downstream application or availability of the database. However, control remains an issue when extracting money and/or talent from the private sector. One possible approach to generate such a model system is to offer the commercial investor a three-month period of exclusive rights that will be followed by a free public-domain period that fosters community and makes the information available to the academic community.

Session #2: Overview of Biomarker Research Networks:

Speakers: **Sudhir Srivastava**, National Cancer Institute, NIH; **Vicki Seyfert-Margolis**, National Institute of Allergy and Infectious Diseases, NIH

Dr. Srivastava discussed the early NCI detection research network (EDRN), an initiative for the establishment of identifying risk factors and early indicators of cancer. The EDRN is a consortium representing about 200 laboratories at 30 institutions across the country designed to provide a linkage between cancer and biomarker discovery and clinical applications. The management of the consortium is provided by NCI in consultation with a steering committee composed of the principal investigators from each site. An independent advisory committee (AC) advises the steering committee (SC) and the NCI regarding recent progress in biomarkers research and suggests avenues for the consortium to consider. A data management and coordinating center (DMCC) manages information flow across the centers and labs.

The two challenges of such an enterprise are how to communicate within the resource labs and how to make the participating labs feel that there is good reason to continue inputting into the resource. To accomplish these goals, the consortium has many subcommittees responsible for formulating the specific goals, plans, and policy for the network. Each subcommittee addresses a specific issue such as the criteria for moving biomarkers from discovery to application or integrating the EDRN discoveries with the larger community. The EDRN also has an Associate Membership Program that allows interested parties to submit proposals for funding, in particular to develop informatics tools for the consortium. The EDRN currently has two web sites http://cancer.gov/edrn a password-protected site accessible to EDRN investigators and a public site for information, news, and contacts for the consortium. In addition, the EDRN has been featured in major journals and conducts annual workshops and conferences to investigate new frontiers in cancer detection and diagnosis research.

The EDRN employs a systematic process for taking a biomarker from development to validation. First, the investigator submits a proposal to the steering committee for review. If approved, the biomarkers validation laboratory conducts an assay cross-check. If the cross-check is approved, study designs and protocols can then be established with assistance of the SC, AC, and DMCC. Studies can then be performed in conjunction with the Clinical Centers Validation Laboratory.

Specific goals of the EDRN include:

- Rapidly identifying and validating promising biomarkers for large scale studies
- Conducting early phases of clinical/epidemiological studies
- Establishing an EDRN informatics linkage with the NCI Enterprise System
- Formulating a molecular taxonomy of pre-cancerous lesions establishing standards for pre-cancer classification
- Establishing standards for analytical and clinical validation of biomarkers

To achieve these goals, the ERDN has divided the bioinformatics tasks into three categories. The first, knowledge engineering, is concerned with the development of algorithms to perform tasks that require expert knowledge such as medical diagnosis and array data analysis. Discussion is currently underway to form a partnership with IBM, as it has proteomics and genomics software under development that needs validation through data. Also, the NCI Director's Challenge investigator-led Analytical Group will be consulted for data mining efforts within the EDRN. The second focus is the integration of knowledge into a Biomarkers Knowledge Base and ultimately into a Biomarkers Knowledge Center. The Biomarkers Knowledge Base will be a resource for hypothesis generation, meta-analysis, information dissemination to avoid duplication, and patient management. NIH is collaborating with the Jet Propulsion Laboratory. In a pilot project for this effort, the EDRN is setting up common data elements, data elements for laboratory assays, and collaborative groups to decipher organ-specific biomarkers. The final category, knowledge representation, focuses on the representation of knowledge, such as molecular taxonomies, to facilitate problem-solving programs. Informatics challenges that the EDRN faces include the ordering of biomarkers data for efficient retrieval, query, and interpretation, data mining tools, sharing of data between various platforms, data heterogeneity, and privacy concerns regarding clinical data.

Dr. Seyfert-Margolis spoke about the NIAID Immune Tolerance Network http://www.immunetolerance.org, a collaborative network of 40 research institutions that addresses clinical trials in kidney and islet transplantation, clinical trials in autoimmune disease, development of tolerance assays, and clinical trials in asthma and allergic diseases. The network is designed to solicit, develop, implement, and assess clinical strategies and biological assays for the purposes of inducing, maintaining, and monitoring tolerance in humans for these conditions. The network encompasses two different components: clinical trials and mechanistic studies in kidney and islet transplantation, autoimmune disease, and asthma and allergic diseases and development and validation of assays to measure the induction, maintenance, and/or loss of immune tolerance in humans. Diseases currently under investigation include Type-1 diabetes, rheumatoid arthritis, multiple sclerosis, and lupus. Core assay facilities currently include a PCR-based gene expression and polymorphisms core, a pharmacogenomics and microarray core, and MHC-peptide complex core, and a cell-based tolerance assay core. Submitted samples are bar-coded, sent to a central repository and then distributed to the various core facilities. The information then returns to a central database. Informatics challenges to this model include tracking the bar-coded information so that each clinical site has access to their samples only via the bar code and constructing web-based clinical case report forms that establish the top 20 criteria that each disease group wishes to collect. In response to Dr. Rochelle Long's query about accessing the stored information, Dr. Seyfert-Margolis noted two strategies. The first is to make available all of the information from the assays to the source clinical site. Ultimately, however, cross-comparative analyses between the different clinical studies will prove more useful, and this is being approached in two ways: either a central steering committee will have access or the network will accept applications for data analysis in which applicants state specifically which data they wish to see. The ultimate goal is to make as much of the information public as is possible, although the network is still addressing this issue.

The network accepts proposals from network and non-network investigators for concept review by the Scientific Review Committee and the Network Steering Committee. Applications are then reviewed and prioritized and resources allocated by the Network Executive Committee and the Budget Committee. Protocol development for regulatory submission is conducted through the Clinical Trials Oversight Committee and the Clinical Trials Coordinating Committee, and the minimum time frame for the entire process is approximately 20 weeks. Because the network conducts clinical studies surrounding a mechanism rather than to test a drug, it awarded one 7-year, $144 million contract focused on human studies to a specific investigator. Dr. Seyfert-Margolis also stressed that the network differs from the EDRN in that the network is empowered to become its own mini-funding agency.

Session #3: Interoperability and Data Architecture for Metadata Development:

Speakers: **Dan Crichton and Steve Hughes**, California Institute of Technology/Jet Propulsion Laboratory/National Aeronautics and Space Administration
Discussants: **Gary Strong**, Defense Advanced Research Projects Agency and National Science Foundation

The speakers discussed the role of enterprise computing at the Jet Propulsion Laboratory (JPL), NASA's lead center for robotic exploration of the solar system, and the importance of establishing an enterprise data architecture for NASA and the JPL. Due to the vast quantity of data acquired from various missions, NASA has sought to design an architecture that will relate the various types of information. Traditionally, the strategy has been to build a solution for a given center and then attempt to make this interoperable with other centers, but difficulties have demanded a paradigm shift toward addressing the issue from a global horizontal view that spans the entire enterprise. The area of knowledge management examines how to capture and manage knowledge across NASA and is supported by an enterprise data architecture (EDA). Key components of EDA are data interoperability, data sharing, data access, and the ability to facilitate access to given data. A typical challenge for EDA construction would be to determine how to construct one interface to access ten separate databases rather than constructing ten separate interfaces.

One of the key points to EDA is to enable the idea of non-discovery and to construct a database that avoids replication of data. Current databases have no standard interface or standard agency-wide meta-language, nor is there a common registry of data products. Thus, a scientist wishing to know whether a data product exists must currently search each individual data system. Moreover, the data are heterogeneous and use different management platforms. Our solution is to build a data architecture to interrelate data across these disparate systems by focusing on metadata management and a framework for interoperability. Metadata represents a classification or identification that allows the user to interpret the data in a useful context. For example, the value "55" is a piece of data, but without knowing whether it refers to age or miles per hour, it is of no use in a data system. One step toward organizing such data is to build metadata repositories that describe currently distributed data products (e.g. location, target, observation date, etc). Standards that the JPL has investigated include ISO/IEC 11179, which provides definitions for how to describe data elements, and Dublin Core, a widely accepted specification for common data elements that exist in every metadata dictionary.

In the computer industry, middleware is a general term for programming that "glues together" or mediates between two separate and already existing programs. Middleware often operates with an electronic data interchange (EDI) mechanism, and it is necessary to encapsulate the data systems away from the user so that the scientists no longer need to understand the topology of how the architecture was built. Thus, middleware can link application, data, and user interfaces while hiding the unique interfaces from the users. Given the pre-existing culture at the JPL and NASA, middleware becomes essential as it is impossible to dictate that scientists use a common standard and re-implement their systems accordingly. The second challenge of developing middleware is to devise a way to exchange data, and we have devised some mechanisms for this using XML. The middleware framework does not rely upon the creation of new technologies but instead uses existing technology to map the client to various sources. Middleware does not allow for analysis of how individual pieces of data relate to each other; rather, it provides the infrastructure to mine the data for knowledge discovery. The Object Oriented Data Technology Task (OODT), funded by the Office of Space Science at NASA, provides a framework for managing data access and interoperability. Three components that we have focused on are:

- Constructing a service to archive data into the overall architecture for long-term data management
- Managing registries of data products currently used in the community through a profile service
- Creating a product server that allows a user query to be mapped into each individual local database

OODT pilot activity includes a partnership with the planetary data system (PDS) to address interoperability across 10 PDS silos, building a generic XML document type definition to support PDS data dictionary and metadata infrastructure, demonstrating how a science query can return data across the PDS nodes, and demonstrating how the same interface can return information between planetary and astrophysics data systems. OODT metadata development has focused on the creation of a metadata registry to mange the semantics of data shared within and between domains. This registry is comprised of a technology base, a data dictionary, an ontology, and XML for communication. XML was chosen as the language for communication because it is language-neutral and allows focusing on the problem of metadata. Furthermore, XML allows the designer to separate the data from the transport mechanism (i.e. Common Object Request Broker Architecture (CORBA) vs. XML-over-CORBA). The PDS experience with the Planetary Science Data Dictionary has shown the criticality of metadata in enabling data sharing and system interoperability.

The PDS, the official planetary science data archive for the NASA Office of Space Science (OSS) Solar System Exploration (SSE), represents a case study for the enterprise data architecture concept. PDS has been in existence for 10 years and is chartered to ensure that SSE planetary data are archived and available to the scientific community. PDS is also a distributed system designed to optimize scientific oversight in the archiving process. Objectives of the PDS include:

- publishing and disseminating documented data sets for use in scientific analysis
- assisting with projects to deign, generate, and validate data products for placement in archives
- developing and maintaining archive data standards to ensure retrievability for 50 years
- providing expert scientific help to the user community

The goal of the PDS archiving system is for each data set to be autonomous. All information required to understand and interpret the data should therefore be included in the archive. To that end, the archive package includes raw data, data calibrated to physical units, calibration data and algorithms, ancillary data, higher level data products, and metadata. It must be stressed, however, that science, and not technology, drives the system.

PDS is structured as a distributed system designed to optimize scientific oversight in the archiving process. Therefore, it is managed by discipline scientists in conjunction with the project manager. The PDS science discipline nodes provide archival of data and supporting documentation, expertise in data interpretation and the design of future observations, and distribution of data to the community. The PDS central node at the JPL oversees program management, project engineering, and standards development. Currently, the PDS has produced a peer-reviewed archive of Solar System Exploration Data, developed a robust standards architecture, and developed a science-driven management structure. However, in spite of the World Wide Web and a common standards architecture, the PDS continues to be a collection of heterogeneous data systems with little resource sharing. The OODT is addressing this issue as follows:

- prototyping a PDS profile service that will manage metadata profiles for data sets, data products, and data systems

- prototyping PDS product servers to integrate individual data systems
- promoting the use of archive services by mission projects for more efficient production of data products

The speakers concluded their talks with an on-line demonstration of the system.

Session #4: Needs Assessment of User Requirements for Research Network Information Management:

Speaker: **Richard Morris**, National Institute of Allergy and Infectious Diseases, NIH

Dr. Morris provided more global perspectives on requirement specifications and the need for establishing a generalized but variable reference of what must be accomplished at a high level when addressing more localized problems. He stressed that his role today is that of the healthy skeptic of biomarkers and the need for a biomarkers knowledge system. The current status of systems development can best be categorized as the "vision" stage in the sense that all of the paperwork demonstrates a powerful vision on the need for a biomarkers knowledge system. However, no two people agree about the expectations of the system. Thus, Dr. Morris argued that we must be very precise at many levels about what we want the knowledge system to do and how it should accomplish these tasks. One of the reasons for the current failure in generating a working system is that the standard operating procedures being developed for the system are not working well.

Dr. Morris cited the recent Firestone tire recall as an example of such a failure and as an analogous situation. He notes that Firestone experienced more than 750 incidents over a 10-year period without responding, and this ultimately led to as many as 62 deaths and forced a recall of 650 million tires. The current situation may be summed as follows[1]: divergent beliefs with respect to the standard operating procedures of tire inflation, 193 probes from regulatory agencies, fines from as many as 16 countries, current Congressional inquiry, and the largest recall since 1982 generates a *Wall Street Journal* headline that essentially states that management "didn't know that we needed a database."

Although the research community is trying to capture mission-critical information, the progression of science makes the mission a moving target. With 5000 articles published per week, how do we craft a knowledge system that keep abreast of the knowledge management requirements? One lesson to be learned from the Ford-Firestone case is that every event has multiple causes, and even though events are spread across time and space, only a subset of the events is actually relevant. The production of tires was only a problem on certain days given flaws in raw materials, etc., so how do we characterize such a heterogeneous data set to avoid the 62 deaths that Ford may be liable for? One message to take away from this experience is the sense of urgency. Could Ford have crafted a knowledge management system to avoid these kinds of tragedies? Although everyone believes in the concept and potential application of biomarkers in clinical studies, we must now address unanswered questions. Although everyone is noting that it is time for a change, a clear and convincing statement has yet to emerge regarding the driver for a new system. Although everyone claims that a better system is needed and that threats are pending if we fail to do so, no one will act unless there is a consequence for inaction. A concern that I have regards the leadership of such an initiative, i.e. is there a balance between the biological and the computational scientists addressing these issues? Although it is indicated that numerous organizations are addressing the problem, there are precious few success stories to inspire investment by decision-makers.

In certain cases, the driver is apparent. For example, in sequencing the human genome, chip technology has resulted in a flood of information. Thus, a technology disruption may be the driving force for a solution, and we have an abundance of tools to address this. The problem is that these tools are point solutions that remain unintegrated with one another. They solve clinical problems rather than basic biological problems and thus lack the robustness to handle some of the downstream information management issues. However, there is a promise for action, for we have the chance to define a systems architecture that is open, evolvable and collaborative over time. We have got an unprecedented opportunity today to increase scientific productivity in the clinical setting and an opportunity to measure the results of the higher system performance in unprecedented ways.

Data representation issues must be clearly addressed to avoid problems with correctness, verifiability, and reusability of data. Data are generated today at an incredible rate, and we must be prepared to make those data more portable and interoperable. Middleware must be developed to handle this data load because there is an incredible threat of fault tolerance in a multi-tiered environment. Although many of the success stories in this arena are present today, most have been working in isolation. What is needed now is to clarify specifically what transactions are most important, what processes need support, and how to operate from this basis. One of the divisive issues in this quest is a cultural difference between the bench scientists and the computer architects. The bench science response to systems specifications involves notions of constraint of exploration, yet in personal experience with databases and languages, I see a robustness and lack of constraint that is not fully appreciated by the research community. By dispelling these myths, we can begin to craft solutions. A framework for doing so will involve the following steps: defining clearly the domain, committing this definition to written documents, addressing performance requirements, archiving that analysis for frequent referral, and letting that provide the vocabulary for the data design. The data design represents the area of overlap between the natural-language description of wants and needs and a machine-readable language. Another strategic approach involves selecting what will be tested and prototyped first, finding a high-priority functionality that is within the various domains spanning a system such as this, prototyping this on a small scale, and disseminating it broadly in the earliest stage. Finally, we must be explicit about the business rules and the measures of performance that will constitute success. In conclusion, the opportunity today is to at least close the gap between the point of discovery and the point of care.

[1] *Wall Street Journal*, August 10, 2000.

Session #5: Standards and Models for Data Sharing and Archiving:

Speakers: **David Christiansen**, Genentech, Inc.; **W. David Benton**, SmithKline Beecham
Discussant: **Randy Levin**, U.S. Food and Drug Administration

Dr. Christiansen discussed the role of the Clinical Data Interchange Standards Consortium (CDISC) in relation to the biomarkers knowledge system. CDISC is an open, multidisciplinary, non-profit organization committed to 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. The current state of the biomarkers knowledge system includes numerous heterogeneous data sources and a wide variety of data users. Therefore, the ideal biomarkers knowledge system must accept data of different quality and inconsistent documentation from an ever-increasing number of sources yet provide access to an ever-increasing number of users wishing to explore unknown relationships using yet-to-be-defined analytical techniques! The standardization of metadata can provide a solution, and regulatory submission standards may be a stating point for a biomarkers knowledge system. Goals of clinical trials interchange standards include:

- A nearly seamless exchange of data across protocols, companies, and compounds
- Effortless archiving of data and metadata for future review or regulatory audit
- Integration of data from a wide variety of applications and systems
- Facilitated reviews of regulatory submissions
- Improvements in data quality

To achieve these ends, CDISC has adopted the following principles:

- Lead the development of standard data models that improve process efficiency while supporting the scientific nature of clinical research
- Recognize the ultimate goal of creating regulatory submissions that allow for flexibility in scientific content yet are easily interpreted, understood, and navigated by regulatory reviewers
- Acknowledge that data content, structure and quality of the data models are independent of implementation strategy and platform
- Maintain a global, multidisciplinary, cross-functional composition for CDISC and its working groups
- Provide educational programs on CDISC standard, models, values and benefits

CDISC currently uses two approaches, submission data modeling (SDM) and operational data modeling (ODM), to address clinical trial data. The SDM metadata approach organizes datasets according to FDA guidelines, such as one case report tabulation dataset for each clinical domain (demographics, vital signs, adverse events, etc). Dataset attributes such as name, description, and file location are defined and additional attributes are added to facilitate knowledge transfer. Common selection variables are then added to all datasets. This models allows reviewers to replicate most analyses with minimal transformations while enabling them to view and subset the data used in any analysis without complex programming. The SDM will have complete metadata models for 12 safety domains by October 2000.

Model requirements for the ODM include developing interchange standards and facilitating data interchange between laboratories, sponsoring companies, and reporting systems. The ODM group is designed to address the following:

- Support the interchange and archiving of data
- Enable the interchange between applications used in collecting, managing, analyzing, and archiving
- Enable the full description of all data and metadata required to produce regulatory submissions
- Reduce accumulation and conversion costs

The ODM Version 1.0 (September 2000) supports the basic interchange between applications, an audit trail, and reconciliation with the Submission Group's model. Current issues for consideration in future versions include a data clarification history, real-time interfaces, and complex "use case" application interoperability. SDM and ODM are linked since SDM is defining content for ODM XML elements such as protocol and items. Furthermore, some SDM metadata will be XML tags in CRF and future submission datasets to facilitate the traceability of CRF data from source to submission.

CDISC has several advisory boards and working groups to facilitate this process. The Testing and Applications (TAP) Group provides a means for testing the SDM and ODM with real data and applying them in specific scenarios. A Testing team is currently being formed, and a team devoted to laboratory data issues was initiated in August 2000. The Education Working Group (EDU) provides educational information and courses on the CDISC standards. The team, plan, and objectives are in formation, and a CD-ROM was provided by Quintiles in August 2000. The Industry Advisory Board, comprised of one representative from each Corporate Sponsor who contributed CDISC "seed" funding, will advise CDISC on strategic planning. A Scientific Advisory Board, which has no fiduciary responsibilities, is currently being formed and will advise CDISC on scientific issues related to standards development. In addition, CDISC has liaisons with the FDA and has collaborated with HL 7 representatives to convert the CDISC ODM Model DTD to HL 7 version 3.0.

Dr. Feng raised the issue of FDA guidelines driving a clear end product, which does not exist in biomarker research. Dr. Christiansen responded by noting that a scientist wishing to analyze a set of data likely wants much of the same information as does an FDA reviewer. Thus, a standardization or metadata will help in each case, as it provides information about the organization of the data set. Dr. Levin added that the key to success with CDISC has been to tackle the problem piecemeal rather than attempting to develop a panacea for all possible problems.

Dr. Benton discussed the Object Management Group (OMG) and the Life Science Research Domain Task Force (LSR DTF) at SmithKline Beecham. He began by focusing on the role of integration in assisting the transformation of data to information and to knowledge. Integration is made difficult in the current environment due to impedance mismatches between organization and development methods and software systems heterogeneity. Dr. Benton cited a recent paper by Duane Truex, et.al. (Communications of the ACM 42(8):117-123; August 1999) that discusses new perspectives required for IT support as businesses change from stable to emergent operating models. Dr. Benton then noted that the current software crisis in life sciences research is due to the data-driven yet software-dependent nature of the disciplines involved. Life sciences research thus occurs in a heterogeneous computational environment, and productivity may therefore be improved by integration, interoperability, and reuse of computational resources and artifacts. The solution does not lie in converting all software to a single language or hardware platform; rather, the research community must collaborate to develop standards for interoperability and cultivate a common marketplace. Thus, we must keep diversity but provide universal interoperability, integration, and flexibility. Enabling technologies for this approach include object-oriented software, common object request broker architecture (CORBA), and software components with industry-standard interfaces.

Dr. Benton stressed thinking of computer architecture as analogous to a set of building blocks. Thus, smaller parts allow for more flexible shapes and have more uniform interface media. The issue then becomes understanding how to organize a distributed application of smaller blocks rather than the number of tiers created from the blocks. Most applications tend to follow a common structural pattern: presentation, analysis, and storage. The key thus becomes focusing on boundaries and interfaces rather than the internal details of how components are constructed, as these details will constantly evolve. The "glue" that binds these interfaces between blocks is the object request broker (ORB). Common ORB Architecture (CORBA) is a set of industry standard specifications for software interfaces and distributed computing based on object technology and provides a medium for integrating various components.

The OMG, founded in 1989, is the world's largest software consortium and is comprised of over 860 companies. The organization is dedicated to creating and popularizing object-oriented standards for distributed application integration based on existing technology. Through an open, consensus-based process, the ORB facilitates the creation of a multi-vendor, competitive/cooperative marketplace of tools and components that are guaranteed to interoperate. The OMG accepts proposals with an eye to a 12 month implementation period once the submission is approved and finalized. The OMG adopts and publishes Interface Specifications chosen from existing products through a competitive selection process, and these Interface Specifications are freely available to members and non-members, although the interface implementations must be available commercially from an OMG corporate member. The OMG LSR Domain Task Force is comprised of representatives from 40 OMG member companies and is charged with adopting CORBA interface specifications to enable interoperable software components in numerous components of life science research. The OMG mission is defined by the participants, and further information is available at http://www.omg.org/homepages/lsr.

Current LSR working groups include architecture and roadmap, bibliographic services, cheminformatics, clinical trials, entity identification, gene expression, macromolecular structure, sequence analysis, visualization and user interfaces, web site, and workflow. The LSR architecture and roadmap group investigates the partitioning of the domain into sub-units and the relationships and interfaces between those units and coordinates the LSR strategic plan for scheduling and prioritization of planned activities. Current LSR technology adoptions include RFPs for biomolecular sequence analysis and genomic maps and revised submissions are under review for a bibliographic query service and a macromolecular structure proposal. Information about the biomolecular sequence analysis project can be found at http://corba.ebi.ac.uk/openBSA. Recent LSR RFPs include an entity identification service, gene expression, and chemical structure access and representation, and forthcoming RFPs and RFIs include a clinical trials laboratory data interchange, chemical synthesis, compound management, extensions to BSA analyses, and gene analysis. The OMG/LSR can serve the biomarkers community in the following applications:

- As a process/organizational sample
- As a source of useful patterns
- As a source of useful specifications
- As a "sponsor" of useful implementations
- As an organizational home for developing interface specifications for distributed objects for biomarkers knowledge systems

Session #6: Data Dictionaries—Common Data Elements:

Speaker: **Clement McDonald**, Regenstrief Institute, Indiana University School of Medicine

Dr. McDonald began by recounting his background in standardizing a medical record system. At a key meeting in 1984, two concepts emerged: there is no sense in standardizing every single aspect of the universe, and that standardization of any small component could require an extensive output of money and time. Fortunately, much has improved in the last 15 years, and many systems currently in place represent the latest evolutions of concepts that initially seemed difficult. Opportunities in health care have driven the creation of standard vocabularies such as LOINC and HL7, although many barriers remain in terms of the quantity and differences in patient data. Local codes for even the most simplistic variables, such as gender, vary widely, and privacy issues continue to loom. Today there is a fundamental difference in the world view between researchers and regulators and source clinical system developers. Researchers and regulators tend to construe data sets as "flat" sets having one value per field. By contrast, the operating system view, which is seen in lab systems, billing systems, pharmacy systems, and electronic medical record systems, represents "stacked" data sets. There is at least one kind of clinical value per record, and the variable definition is carried in another master file. This allows the user to change or add attributes without having to rebuild the database while allowing the record to retain other information about the data such as who collected the data, the data range, and when the data were delivered. In the past, financial considerations and storage space promoted the collapsing of data, although the amount of data should not be a concern today.

Currently available products to improve the standardization of data include the HL standard as well as standardized vocabularies such as LOINC, SNOMED, NDC-2 and NDC-prime. Languages such as XML are also available, but they are not panaceas, as one message may be represented in multiple ways using the language. HL7, the dominant code for clinical data messaging, is currently used for a variety of applications including registration, laboratory, clinical results, and orders. It has additional capabilities to send real time updates and corrections. LOINC, by contrast, is a database of observations or variables. These are Internet-available, which makes them both convenient and free. Furthermore, LOINC has been translated into other languages including German, French, and Italian. There is also a program that accompanies LOINC to assist mapping into LOINC.

One participant inquired whether HL7 was a national or an international standard, and, although the exact definition of "international" has not been established in this case, HL7 continues to spread in terms of numbers of users. In response to a query regarding the persons responsible for the development of LOINC, Dr. McDonald stated that the concept began with representatives from seven laboratories and continues to grow as new experts participate. One advantage to the approach is that it is finite in its scope. From the perspective of research organizations wishing to develop their own databases, it is important to consider that LOINC currently contains only variables reported at commercial grade levels and not their variants. The first step to addressing the biomarkers issue is to formulate an initial catalog of enumeration of the names of the variables without respect to their measurement. It will also be advantageous to work toward having a catalog of all of the nomenclature for the various methods currently in use. In response to a question about histologic specimens, their interpretation and the report coding, Dr. McDonald noted that SNOMED has much of the rich vocabulary that is necessary for such a project although their current pricing and use policies are prohibitive. If an agreement can be worked out such that pricing can be controlled for all users, then SNOMED may represent a base for the final answer.

Summary of Directions for Data Architecture and Knowledge Environment Development for Biomarker Research (all participants):

The final segment of the meeting consisted of a round-table discussion of the suggestions and information provided throughout the meeting as well as an opportunity for other speakers to present ideas and comments. The discussion began with a brief presentation by Ken Buetow from the Center for Bioinformatics at the National Cancer Institute (NCI). Dr. Buetow described the NCI initiative designed to address bioinformatics challenges and integrate disparate bioinformatics domains. To achieve these goals, the Center for Bioinformatics currently integrates four core units of research interest: the Cancer Genome Anatomy Project, the Mouse Models for Human Cancer program, a clinical trials division, and the Director's Challenge, which investigates molecular signatures of disease states and phenotypes. The Center for Bioinformatics provides a central infrastructure to bring these units together, and individuals associated with each core bring their own expertise to the initiative. The Center for Bioinformatics aims to provide an information transfer infrastructure, to examine various models for the distribution of information and knowledge, and to standardize the vocabulary when necessary for optimal communication. In order to bridge the gap between the research laboratory and the information architecture, a user support component has been budgeted into this model in order to facilitate training of laboratory members in the computer skills necessary for data storage and retrieval.

Dr. John Hewes from the National Institute of Standards and Technology then commented on the future of implementing combinatorial methods into bioinformatics by drawing a parallel between the problems encountered in the life sciences and similar situations in engineering disciplines. In particular, he expressed concern that many people are solving specific problems such as interoperability without considering solutions that extend beyond bioinformatics. Dr. Hewes then proposed the paradigm of meta-information rather than metadata and stressed examining how the convergence of disciplines can actually create new technologies. He noted that it is important to consider how to track the movement of data points over time, as this approach can generate valuable insight into the construction of a workable information architecture.

Dr. Silvia Spengler, Program Officer for Biological Databases and Informatics at the National Science Foundation, then followed up on the comments proposed earlier by Dr. Long. She stressed the importance of considering three factors when designing a knowledge system: the users for a given informatics project, ontologies, and statistical validity, or how to envision the linkage between particular biomarkers and their associated clinical endpoints. This final point was seconded by Dr. Downing, who noted that statistical components are the keys to the validation process for proposed biomarkers.

One key component that was agreed upon by all participants was the need to address local, less global problems initially before trying to solve more complex and universal information architecture issues. Dr. Clement McDonald suggested to begin by considering how to record only the information that you as a researcher are interesting in recording. Then, consider what should be characterized about these observations. Often, many of the larger issues revolve around the collection of data at the time of submission rather than deficiencies in the science involved. Several participants also stressed the need for the communities involved to agree on common definitions and to establish universal guidelines for clinical trials as first steps in building a working knowledge environment. Dr. Krishan Arora from the PhRMA Foundation suggested that the answers to the global problems may lie in simple solutions analogous to variable files or bar codes.

Much discussion then centered on the psychology of the researchers and the culture of the research environment. In particular, issues of control and release of data must take into account the sociology of the scientific community. Questions raised included:

- How do we get investigators involved in the projects to relinquish control of their data?
- What motivates researchers to deposit their collected data in electronic form?
- What role do publishers have in the charge of information architecture?

In response to such questions, several participants stressed the differences between technical challenges and cultural obstacles. As many of the fields currently generating data are new and evolving, it is easy for a researcher to become ingrained within a particular paradigm. In order to combat this tendency, the solution may be to pick a smaller, more workable problem as a starting point before tackling larger issues. Dr. Michael Gilson of the University of Maryland suggested that the important concept was to capture as much data as possible in databases. He noted that data saved in a database have added value because they can outlive interpretation. Several ways to promote this data capture are to provide motivators (perhaps a job for funding agencies) and to reach community consensus on data that can be entered into any database rather than into a database accessible by a select audience. However, cultural limitations may still affect the implementation of such a plan, and Dr. Atul Butte of the Harvard University Medical School noted that the annotations make useful the raw microarray data generated in genomics research.

In response to the final question regarding the role of publishers, Dr. Bradford stressed that the strength of a publisher lies in promotion rather than information architecture. She noted that journals often wait for the scientific community to signal particular consensus specifications, yet the research community often expects the journals to perform this task. *Science*, in particular, has been approached by several projects that have pushed for including all data within a certain database. Although the NCI-OUP's CRADA and the AAAS' STKE systems have evolved from print resources, most print media do not have the bioinformatics infrastructure to oversee information architecture for universal databases.

Finally, participants debated the nature of the information systems under consideration. Dr. Rajeev Gopal of Hughes Network Systems noted that a distinction must be made between the scientific and engineering processes. He proposed to focus on the engineering process, even though it can potentially limit the freedom of expression in the non-structured information of the scientific community. To address this dilemma, he suggested a strategy of building the information architecture requirements and then re-building as necessary in order to promote the maximum amount of freedom yet obtain a workable system. Dr. Butte disagreed with this approach on the premise that biological science is fundamentally a discovery science in which the investigator does not always know about what he or she should be considering as the discovery process unfolds. A balance between the two must be achieved, however, in order to craft a working solution. Dr. McDonald responded by noting that the situation represents a biologic principle but an engineering problem. Dr. Robert Robbins of the Early Detection Research Network noted that information retrieval is often a question of subject matter, and he offered several possible solutions: 1) the data resource could contain both positive and negative findings and observations, thus requiring an open information retrieval system such as an inference engine, or 2) the data resource could contain only pure facts, thus making it more of a closed system. What is missing is the support structure that bridges these two types of approaches; as a result, researchers lack the real keys necessary to associate larger areas, such as gene data with protein number.

Dr. Tom Lewis then offered some closing thoughts on the day's discourse. He stressed the need to put energies into smaller-scale problems initially. Furthermore, he noted that we must consider methodologies for developing architecture for open-ended environments. This task will require a core of motivated individuals who enjoy both the clinical aspects and the bioinformatics issues associated with biomarkers. This core group would then be charged with examining various models and designing targeted experiments to assess the strengths and weaknesses of the specific models before proposing large-scale application.

As it was difficult for the participants to agree on one unified approach for attacking the problems associated with developing a bioinformatics knowledge system for biomarkers, Dr. Butte commented on the fact that technology has allowed researchers to generate biomarkers rapidly. As a result, it would be helpful to view the current bioinformatics issues not simply as problems but as outcroppings from a meaningful basis of research.

OFFICE OF SCIENCE POLICY
OFFICE OF THE DIRECTOR
NATIONAL INSTITUTES OF HEALTH

Biomarkers Knowledge System

Bethesda Marriott
5151 Pooks Hill Road
Bethesda, MD

September 8, 2000

AGENDA

7:30 a.m. - 8:15 a.m. **Continental Breakfast**

8:15 a.m. - 8:25 a.m. **Welcome**
Lana R. Skirboll, Ph.D., Director, Office of Science Policy, National Institutes of Health

8:25 a.m. - 8:35 a.m. **Introductions**
All participants

8:35 a.m. - 8:40 a.m. **Goals and Objectives**
Gregory J. Downing, D.O., Ph.D., Office of Science Policy and Planning, National Institutes of Health

8:40 a.m. - 10:20 a.m. **Knowledge Environments—Lessons Learned, New Directions**

Cancer Spectrum
Michael Stout, M.Sc., C.S., Oxford University Press

Signal Transduction Knowledge Environment
Monica M. Bradford, American Association for the Advancement of Science

Discussants
Rochelle M. Long, Ph.D., National Institute of General Medical Sciences
Stephen M. Maurer, J.D., University of California–Berkeley

Open Discussion

10:20 a.m. - 10:40 a.m. **Break**

10:40 a.m. - 11:15 a.m. **Overview of Biomarker Research Networks**

Early Detection Research Network
Sudhir Srivastava, Ph.D., M.P.H., National Cancer Institute, National Institutes of Health

	Immune Tolerance Network *Vicki L. Seyfert-Margolis, Ph.D., National Institute of Allergy and Infectious Diseases, National Institutes of Health*
11:15 a.m. - 12:00 noon	**Interoperability and Data Architecture for Metadata Development** *Dan Crichton, M.S., California Institute of Technology/Jet Propulsion Laboratory/National Aeronautics and Space Administration* Discussant *Gary Strong, Ph.D., M.S.E.E., Defense Advanced Research Projects Agency* Open Discussion
12:00 noon - 12:30 p.m.	**Needs Assessment of User Requirements for Research Network Information Management** *Richard W. Morris, Ph.D., M.S.E., National Institute of Allergy and Infectious Diseases, National Institutes of Health*
12:30 p.m. - 1:00 p.m.	**Working Lunch**
1:00 p.m. - 2:15 p.m.	**Standards and Models for Data Sharing and Archiving** Clinical Data Interchange Standards Consortium *David Christiansen, Dr.P.H., M.B.A., Genentech, Inc.* Discussant *Randy Levin, M.D., U.S. Food and Drug Administration* Object Management Group *W. David Benton, Ph.D., SmithKline Beecham*
2:15 p.m. - 3:00 p.m.	**Data Dictionaries—Common Data Elements** *Clement J. McDonald, M.D., Regenstrief Institute, Indiana University School of Medicine* Open Discussion
3:00 p.m. - 3:15 p.m.	**Break**
3:15 p.m. - 4:25 p.m.	**Summary of Directions for Data Architecture and Knowledge Environment Development for Biomarker Research** *All participants*
4:25 p.m. - 4:30 p.m.	**Closing Remarks** *Gregory J. Downing, D.O., Ph.D., Office of Science Policy and Planning, National Institutes of Health*

OFFICE OF SCIENCE POLICY
OFFICE OF THE DIRECTOR
NATIONAL INSTITUTES OF HEALTH

Biomarkers Knowledge System

Bethesda Marriott
5151 Pooks Hill Road
Bethesda, MD

September 8, 2000

PARTICIPANT LIST

Krishan K. Arora, Ph.D.
Vice President
Research and Development Management Information
Pharmacia
100 Route 206, North
Peapack, NJ 07977
(908) 901-8616
(908) 901-1882 FAX
krishan.arora@am.pnu.com
Representing: PhRMA

W. David Benton, Ph.D.
Director
Intelligent Information Systems
SmithKline Beecham
MS UW2318
709 Swedland Road
King of Prussia, PA 19406
(610) 270-6864
(610) 270-4388 FAX
w_david_benton@sbphrd.com
Representing: Object Management Group

Monica M. Bradford
Managing Editor
Science
American Association for the Advancement
of Science
Room 1051
1200 New York Avenue, NW
Washington, DC 20005
(202) 326-6502
(202) 289-7562 FAX
mbradfor@aaas.org
Representing: Signal Transduction Knowledge Environment

Kenneth H. Buetow, Ph.D.
Director
Center for Bioinformatics
National Cancer Institute
National Institutes of Health
MSC 8302
8424 Helgerman Court
Bethesda, MD 20892-8302
(301) 435-1520
(301) 435-9325 FAX
buetowk@pop.nci.nih.gov

Atul J. Butte, M.D.
Endocrinologist/Bioinformatician
Children's Hospital Informatics Program
Harvard University Medical School
333 Longwood Avenue
Boston, MA 02115
(617) 355-2561
(801) 729-7231 FAX
atul_butte@harvard.edu

Kristen Chambers, M.S.
Director of Bioinformatics
Dartmouth Medical School
Evergreen Suite 301
Lebanon, NH 03766
(603) 650-3402
(603) 650-3411 FAX
kristen.chambers@dartmouth.edu
Representing: Early Detection Research Network

David Christiansen, Dr.PH., M.B.A.
Principal Biostatistician
Genentech, Inc.
1 DNA Way
San Francisco, CA 94080-4990
(650) 225-1738
(650) 225-3233 FAX
davec@gene.com
Representing: Clinical Data Interchange Standards Consortium

Joseph F. Contrera, Ph.D.
Director
Regulatory Research and Analysis
Office of Research and Testing
Center for Drug Evaluation and Research
U.S. Food and Drug Administration
HFD-901
5600 Fishers Lane
Rockville, MD 20857
(301) 827-5188
(301) 827-3787 FAX
contrerajf@cder.fda.gov

Dan Crichton, M.S.
Project Element Manager
Enterprise Data Architecture Task
Principal Investigator
Object Oriented Data Technology
Science Data Management and Archiving Section
Jet Propulsion Laboratory
National Aeronautics and Space Administration
California Institute of Technology
M/S 171-264
4800 Oak Grove Drive
Pasadena, CA 91109
(818) 354-9155
(818) 393-3405 FAX
dan.crichton@jpl.nasa.gov

Joel Dobbs, Ph.D.
Vice President of Research Information Systems
Schering-Plough Research Institute
2000 Galloping Hill Road
Kenilworth, NJ 07733
(908) 740-2355
(908) 740-2814 FAX
joel.dobbs@spcorp.com
Representing: PhRMA

Gregory J. Downing, D.O., Ph.D.
Health Science Policy Analyst
Office of Science Policy and Planning
Office of the Director
National Institutes of Health
Building 1, Room 218
9000 Rockville Pike
Bethesda, MD 20892
(301) 594-7740
(301) 402-0280 FAX
downingg@od.nih.gov

Ziding Feng, Ph.D.
Principal Investigator
Data Management and Coordinating Center
Early Detection Research Network
Fred Hutchinson Cancer Research Center
MP-859
1100 Fairview Avenue, North
Seattle, WA 98109-1024
(206) 667-6038
(206) 667-5965 FAX
zfeng@fhcrc.org
Representing: Early Detection Research Network

Michael K. Gilson, M.D., Ph.D.
Associate Professor
Center for Advanced Research in Biotechnology
University of Maryland
9600 Gudelsky Drive
Rockville, MD 20850
(301) 738-6217
(301) 738-6255 FAX
gilson@umbi.umd.edu

Charles A. Goldthwaite, Jr.
Science and Medical Writer
124 B Thomas Drive
Charlottesville, VA 22903
(804) 293-2907
(804) 293-2907 FAX
cag5c@cms.mail.virginia.edu

Rajeev Gopal, Ph.D.
Senior Director
Spaceway Engineering
Hughes Network Systems
11717 Exploration Lane
Germantown, MD 20876
(301) 428-5551
(301) 428-5575 FAX
rgopal@hns.com

Peter Greenwald, M.D., Ph.D.
Director
Division of Cancer Prevention
National Cancer Institute
National Institutes of Health
Building 31, Room 10A-52
MSC 2580
31 Center Drive
Bethesda, MD 20892-2580
(301) 496-6616
(301) 496-9931 FAX
pg37g@nih.gov
Representing: Early Detection Research Network

Demian Harvill, M.S.
Stanford University
Galvez Hall, Room 482
Stanford, CA 94305-6004
(650) 725-6142
(650) 725-9335 FAX
harvill@stanford.edu
Representing: HighWire Press

John D. Hewes, Ph.D., M.S.M.O.T.
Program Manager
Advanced Technology Program
National Institute of Standards and Technology
U.S. Department of Commerce
Administrative Building 101, Room A235
Mail Stop 4730
100 Bureau Drive
Gaithersburg, MD 20899-4730
(301) 975-5416
(301) 548-1087 FAX
john.hewes@nist.gov

J. Steven Hughes, M.S.
Systems Engineer
Advanced Concept Engineering
Science Data Management and Archiving Section
Jet Propulsion Laboratory
National Aeronautics and Space Administration
California Institute of Technology
MS 171-264
4800 Oak Grove Drive
Pasadena, CA 91109-8099
(818) 354-9338
(818) 393-3405 FAX
steve.hughes@jpl.nasa.gov

Robin I. Kawazoe
Director
Office of Science Policy and Planning
Office of the Director
National Institutes of Health
Building 1, Room 218
9000 Rockville Pike
Bethesda, MD 20892
(301) 496-1454
(301) 402-0280 FAX
rk52i@nih.gov

Heather Kincaid
Information Technology and Database Manager
Data Management and Coordinating Center
Early Detection Research Network
Fred Hutchinson Cancer Research Center
MP-859
1100 Fairview Avenue, North
Seattle, WA 98109-1024
(206) 667-7909
(206) 667-5964 FAX
hkincaid@fhcrc.org
Representing: Early Detection Research Network

Barry Kramer, M.D.
Editor
Journal of the National Cancer Institute
Director
Office of Medical Applications of Research
Office of the Director
National Institutes of Health
Building 31, Room 1B-03
MSC 2580
9000 Rockville Pike
Bethesda, MD 20892-2580
(301) 496-5641
(301) 402-0420 FAX
kramerb@od.nih.gov
Representing: Early Detection Research Network

Randy Levin, M.D.
Associate Director for Electronic Review
Center for Drug Evaluation and Research
U.S. Food and Drug Administration
1451 Rockville Pike
Rockville, MD 20852
(301) 594-5411
(301) 594-6197 FAX
levinr@cder.fda.gov

Thomas L. Lewis, M.D.
Clinical Informatics Consultant
Former Associate Director for Information Systems
Warren G. Magnuson Clinical Center
National Institutes of Health
11801 Gainsborough Road
Potomac, MD 20854-3355
(301) 299-2045
tomlewis@post.harvard.edu

Ronald Lieberman, M.D.
Program Director
Prostate and Urologic Cancer Research Group
Division of Cancer Prevention
National Cancer Institute
National Institutes of Health
Suite 201
Executive Plaza North
6130 Executive Boulevard
Rockville, MD 20852
(301) 594-0456
(301) 402-0553 FAX
rl39r@nih.gov

Peter Y. Lin
Web Developer
Data Management and Coordinating Center
Early Detection Research Network
Fred Hutchinson Cancer Research Center
MP-859
1100 Fairview Avenue, North
Seattle, WA 98109-1024
(206) 667-7335
(206) 667-5964 FAX
plin@fhcrc.org
Representing: Early Detection Research Network

Rochelle M. Long, Ph.D.
Chief
Pharmacological and Physiological Sciences Branch
Pharmacology, Physiology, and Biological
Chemistry Division
National Institute of General Medical Sciences
National Institutes of Health
Building 45, Room 2AS-49A
MSC 6200
45 Center Drive
Bethesda, MD 20892-6200
(301) 594-1826
(301) 480-2802 FAX
longr@nigms.nih.gov

Robert L. Martino, Ph.D.
Acting Scientific Director
Center for Information Technology
Division of Computational Bioscience
National Institutes of Health
Building 12A, Room 2033
MSC 5624
12 South Drive
Bethesda, MD 20892-5624
(301) 496-1112
(301) 402-2867 FAX
robert.martino@nih.gov

Stephen M. Maurer, J.D.
University of California–Berkeley
2632 Hilgard Avenue
Berkeley, CA 94709
(510) 848-3593
(510) 643-9657 FAX
maurer@econ.berkeley.edu
Representing: Human Genome Organization
Mutation Database Initiative

Clement J. McDonald, M.D.
Director
Regenstrief Institute
School of Medicine
Indiana University
RG-5
1050 Wishard Boulevard
Indianapolis, IN 46202
(317) 630-7070
(317) 630-6962 FAX
clem@regen.rg.iupui.edu
Representing: Regenstrief Institute

Johanna McEntyre, Ph.D.
Visiting Research Fellow
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Building 38A, Room 8S806
8600 Rockville Pike
Bethesda, MD 20894
(301) 435-5987
(301) 480-9241 FAX
mcentyre@ncbi.nlm.nih.gov

Richard W. Morris, Ph.D., M.S.E.
Special Expert
Office of Innovative Scientific Research
 Technologies
National Institute of Allergy and Infectious Diseases
National Institutes of Health
Room 5126
MSC 7640
6700-B Rockledge Drive
Rockville, MD 20892-7640
(301) 594-7634
(301) 402-2571 FAX
rmorris@niaid.nih.gov

John Nelson, Ph.D.
Associate Editor
Signal Transduction Knowledge Environment
American Association for the Advancement
 of Science
Room 1028-A
1200 New York Avenue, NW
Washington, DC 20005
(202) 326-8946
(202) 289-7562 FAX
jnelson@aaas.org
Representing: Signal Transduction Knowledge Environment

Daniel Normolle, Ph.D.
Senior Research Associate
Biostatistics Unit
Assistant Research Scientist
Department of Radiation Oncology
Comprehensive Cancer Center
University of Michigan
Medical Inn Building, Room C342
1500 East Medical Center Drive
Ann Arbor, MI 48109-0848
(734) 764-2473
(734) 763-6236 FAX
monk@umich.edu
Representing: Early Detection Research Network

James T. Renfrow, Ph.D.
Chief Information Officer
Manager of Institutional Computing and
 Information Services
Jet Propulsion Laboratory
National Aeronautics and Space Administration
California Institute of Technology
Building 202, Room 204
4800 Oak Grove Drive
Pasadena, CA 91109-8099
(818) 354-9157
(818) 393-1539 FAX
james.t.renfrow@jpl.nasa.gov

Robert J. Robbins, Ph.D.
Co-Investigator
Data Management and Coordinating Center
Early Detection Research Network
Fred Hutchinson Cancer Research Center
LM-120
1100 Fairview Avenue, North
Seattle, WA 98109-1024
(206) 667-4778
(206) 667-2294 FAX
rrobbins@fhcrc.org
Representing: Early Detection Research Network

Vicki L. Seyfert-Margolis, Ph.D
Director
Office of Innovative Scientific Research
 Technologies
National Institute of Allergy and Infectious Diseases
National Institutes of Health
Room 5125
MSC 7640
6700-B Rockledge Drive
Bethesda, MD 20892-7640
(301) 594-7478
(301) 402-2571 FAX
vseyfert@niaid.nih.gov

Lana R. Skirboll, Ph.D.
Director
Office of Science Policy
Office of the Director
National Institutes of Health
Building 1, Room 103
MSC 0143
9000 Rockville Pike
Bethesda, MD 20892-0143
(301) 496-2122
(301) 402-1759 FAX
lana_skirboll@nih.gov

Eric Snowdeal
Senior Informatics Software Engineer
Global Software Group
Motorola, Inc.
IL 94, Second Floor
50 NW Point Drive
Elk Grove Village, IL 60007
(847) 907-8719
(847) 907-8790 FAX
aes022@email.mot.com

Sylvia Spengler, Ph.D.
Program Officer
Biological Databases and Informatics
National Science Foundation
4210 Wilson Boulevard
Arlington, VA 22230
(703) 292-8470
(703) 292-9063 FAX
sspengle@nsf.gov

Sudhir Srivastava, Ph.D., M.P.H.
Chief
Cancer Biomarkers Research Group
Division of Cancer Prevention
National Cancer Institute
National Institutes of Health
Executive Plaza North, Room 330-F
MSC 7346
6130 Executive Boulevard
Bethesda, MD 20892-7346
(301) 496-3983
(301) 402-0816 FAX
ss1a@nih.gov
Representing: Early Detection Research Network

Michael Stout, M.Sc., C.S.
Electronic Development Manager
Academic Division
Electronic Journals Department
Oxford University Press
Great Clarendon Street
Oxford OX2 6DP
United Kingdom
01-865-267695
01-865-267985 FAX
stoutm@oup.co.uk
Representing: Journal of the National Cancer Institute

Gary Strong, Ph.D., M.S.E.E.
Program Manager
Information Technology Office Chair
Human Computer Interaction and Information Management Coordinating Group
Office of Science Technology Policy
Defense Advanced Research Projects Agency
3701 Fairfax Drive
Arlington, VA 22203
(703) 696-2259
(703) 696-4534 FAX
gstrong@darpa.mil
Program Manager
Information Technology Research Experimental and Integrative Activity
National Science Foundation
4201 Wilson Boulevard, Suite 1115
Arlington, VA 22230
(703) 292-8980
(703) 292-9030 FAX
gstrong@nsf.gov

Mark Thornquist, Ph.D.
Co-Principal Investigator
Data Management and Coordinating Center
Early Detection Research Network
Fred Hutchinson Cancer Research Center
MP-859
1100 Fairview Avenue, North
Seattle, WA 98109-1024
(206) 667-2931
(206) 667-5964 FAX
mthornqu@fhcrc.org
Representing: Early Detection Research Network

Thuy Tran
Senior Software Engineer
Jet Propulsion Laboratory
National Aeronautics and Space Administration
California Institute of Technology
MS 126-106
4800 Oak Grove Drive
Pasadena, CA 91101
(626) 844-0620, ext. 226
(626) 844-0638 FAX
thuy.tran@jpl.nasa.gov

Paul F. Uhlir, J.D.
Director
International Scientific and Technical
Information Programs
National Research Council
National Academy of Sciences
2101 Constitution Avenue, NW
Washington, DC 20418
(202) 334-2807
(202) 334-2231 FAX
puhlir@nas.edu