Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2004 Proceedings

Americas Conference on Information Systems (AMCIS)

December 2004

Heterogeneous Data in Federated Networks: A Framework for Solution Development

Deborah Beranek-Lafky
University of California, Irvine

Follow this and additional works at: http://aisel.aisnet.org/amcis2004

Recommended Citation

Beranek-Lafky, Deborah, "Heterogeneous Data in Federated Networks: A Framework for Solution Development" (2004). *AMCIS 2004 Proceedings*. 44.
http://aisel.aisnet.org/amcis2004/44

This material is brought to you by the Americas Conference on Information Systems (AMCIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in AMCIS 2004 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Heterogeneous Data in Federated Networks: A Framework for Solution Development

Deborah Beranek Lafky
University of California, Irvine
dlafky@uci.edu

ABSTRACT

The National Institutes of Health has articulated the need for facilitative technologies aimed at increasing the productivity and efficiency of research networks. There are many organizational issues that must be addressed, but among the most important is informatics. Data heterogeneity among clinical research network nodes is an impediment to achieving knowledge integration. Existing research networks display a diversity of informatics implementations among their nodes. This paper proposes methods for overcoming this diversity with autonomous, adaptive data discovery tools. Grounding this discussion is an examination of a specific research network and an overview of the sources of data heterogeneity within it. A general description of the existing informatics system and how heterogeneity is currently dealt with follows. A less labor-intensive approach, using machine learning concepts within a framework of a domain-specific ontology and controlled vocabulary, is then outlined.

Keywords

Data heterogeneity, clinical research network, genetic epidemiology, ontology, common data element

INTRODUCTION

Scientific Research Strategies in Transition

The size and complexity of inquiries in the bioscience domain have increased tremendously in the past ten years. This trend has been accelerated by the success of the genomics enterprise, which has made available to researchers vast quantities of highly relevant data. The initial timeline to completion of the human genome map was 15 years, with an optimistic completion date sometime in 2005. Halfway through that timeline, the project was only 3% complete. A combination of factors, including research collaboration, as well as competitive pressure and the development of new technologies, resulted in completion of the draft sequence in 2000—five years ahead of schedule (Nass and Stillman, 2003).

The Human Genome Project (HGP) represented a qualitatively different way of doing science and ushered in a new era of "discovery science," characterized by Leroy Hood, father of gene sequencing technology, as defining all the elements of an object and creating from those elements new knowledge independent of hypotheses (Zacks, 2001). The goal of the HGP was simply to discover the sequence of the human genome, without any specific or predetermined hypotheses about it. One result of the HGP was an enormous increase in raw data being made available to science. In the post-genomic era, the pace of data acquisition is even faster.

When presented with an overwhelming quantity of data, the very ability to form a hypothesis can be overpowered. Early predictions were that *status quo* science would adapt and soon develop the ability to channel the torrent of data into a hypothesis-driven model (Gilbert, 1991). Instead, these large and complex problems have pushed the traditional single-investigator hypothesis-driven model of biomedical research to its natural limits. To increase the scope of effort and to be able to address effectively these large complex problems, the clinical research community is beginning to move toward a collaborative model. Taking the success of the HGP, which was partly attributable to new levels of collaboration, as a lesson and a model, the National Institutes of Health (NIH) announced, in 2003, a "Roadmap" for funding future research (Zerhouni, 2003). With the adoption of the Roadmap, the NIH places extraordinary emphasis on "'revolutionary methods of research' focused on scientific questions too complex to be addressed by the single-investigator scientific approach" (Nass and Stillman, 2003). This NIH initiative seeks to expand on an existing array of research networks formed since the mid-1990s, including several within the National Cancer Institute (NCI). Some of these first, experimental, research networks include the Cancer Family Registries (CFR), the Cancer Genetics Network (CGN), and the Early Detection Research Network (EDRN).

Clinical research collaboratives are typically operationalized as federated networks organized around a set of specific aims, producing the Clinical Research Network (CRN). Each network node is represented by a research center led by its own

Principal Investigator. The network is organized in a star topology, with an informatics center at its hub. Each node is independent of the others, but the data that each contributes to the network creates a central repository available to investigators. Because informatics is the main linkage among the CRN nodes, its role is central to the capability of the network to function.

Informatics in the Collaborative Clinical Research Enterprise

Whereas hypothesis-driven science collects only data expected to be relevant to the support of the hypothesis, discovery science seeks emergent knowledge from nonselectively collected data. It follows that specialized informatics structures are needed to acquire, manage, analyze, and disseminate this type and volume of data, since it is quite different in character from previous biology data sets. As the NIH moves forward with the Roadmap, and collaborative research becomes the norm, it will be important for the collaboratives to have available off-the-shelf informatics tools that are specifically designed to support this style of research. There is no such toolkit currently available and, consequently, each new collaborative has had to devise its own informatics tools. This situation unnecessarily slows the pace of research.

This paper discusses informatics issues that arise at the convergence of the collaborative research and discovery science models. Some possible strategies for managing data in this new environment are examined, but the main focus is on an integrational approach, as opposed to a centralized or distributed approach. The integrational approach is to develop a set of tools that are capable of detecting, extracting, and standardizing locally collected data from each node of a research network. The main goals of the integrational approach are: 1) to permit local management of locally-collected data, 2) to facilitate collaborative access to locally-held data repositories, 3) to reduce or eliminate reliance on fragile data mapping schemes, 4) to reduce to a minimum the labor required by each node to participate in the collaborative data resource. The integrational approach may use a central data repository, but the ideal result would be a virtual repository. The integrational approach described in this paper uses ontologies, controlled vocabularies, and intelligent agents to create structured and standardized views into the contributory databases.

EXISTING APPROACHES TO COLLABORATIVE INFORMATICS

The National Cancer Institute (NCI) of the NIH was an early adopter of the research collaboration network structure. Cancer is a complex, multi-faceted problem (Seminara, 1999). Cancer causes involve interactions among environmental factors, infectious agents, genes, and biological pathways. Because of its complexity and diversity, cancer is less susceptible than many problems to a single-hypothesis mode of inquiry and may be better suited to a discovery science research paradigm. Recognizing this, the NCI began, in the mid-1990s, to establish collaborative research networks in several areas. Among these were the Early Detection Research Network (EDRN), which consists of 18 biomarker development laboratories, 3 biomarker validation laboratories, and 9 clinical epidemiology centers. Its goal is to conduct research on molecular, genetic, and other biomarkers that may be useful in early cancer detection and risk assessment (EDRN). The Cancer Family Registries (CFR) is a CRN that consists of 12 clinical and epidemiological centers that are divided into two groups for the study of breast cancer and colon cancer. The goal of the CFR is "to facilitate interdisciplinary studies in the genetic epidemiology of cancer and to provide a flexible, comprehensive and collaborative research infrastructure (CFR, 1996). Another NCI network, and the one that will provide the example for this paper, is the Cancer Genetics Network. The CGN is made up of 8 centers who collaborate in the study of inherited predisposition to cancer by collecting data on cancer in sets of family members. Among these three collaborative networks, there are two basic approaches to data resource sharing. The CGN and CFR employ a centralized approach, which will be discussed in more detail in a later section of this paper. The EDRN initially used a mapping process to map local data models through the use of Common Data Elements (CDEs) and brought it into a common data architecture at the enterprise level. In practice, this has resulted in a central repository of curated data which is fenced off into partitions according to its source (Crichton et al., 2001). This is an intermediate solution between the centralized and the fully integrational approaches. Current efforts within EDRN are focused on abstracting this model out to a virtual repository, which is much more similar to the fully integrational approach (Kincaid et al., 2003).

Data Heterogeneity

In order to understand the requirements for designing a means to integrate disparate data sources, it is helpful to look at the details of those sources. This paper will focus on the Cancer Genetics Network centers to illustrate the nature and extent of data heterogeneity among the network nodes in the context of a previously described data heterogeneity schema (Lafky, 2003). Data heterogeneity within a research network can be viewed as a function of

1) the original purpose of the contributory database,

2) the extent to which the data model of the contributory database does or does not use standardized design elements, e.g. ontologies and/or controlled vocabularies,

3) constraints imposed by external forces, such as choice of database management system and implementation hardware systems.

This heterogeneity can be expressed at several levels. Figure 1 illustrates in a hierarchical view the sources and types of heterogeneity that can exist within data sharing collaboratives. At each level of aggregation, from attribute to knowledge domain, design choices can contribute to the accrual of heterogeneity. As will be discussed in a later section of this paper, an ontology is useful in mitigating this (Musen, 1992). On the formal level, differences in expression and structure can add heterogeneity within the collaborative network. This type of heterogeneity can be addressed through the use of controlled vocabularies and CDEs (Covitz et al., 2003) which reconcile differences in meaning, naming, formatting, structure, and content boundaries.

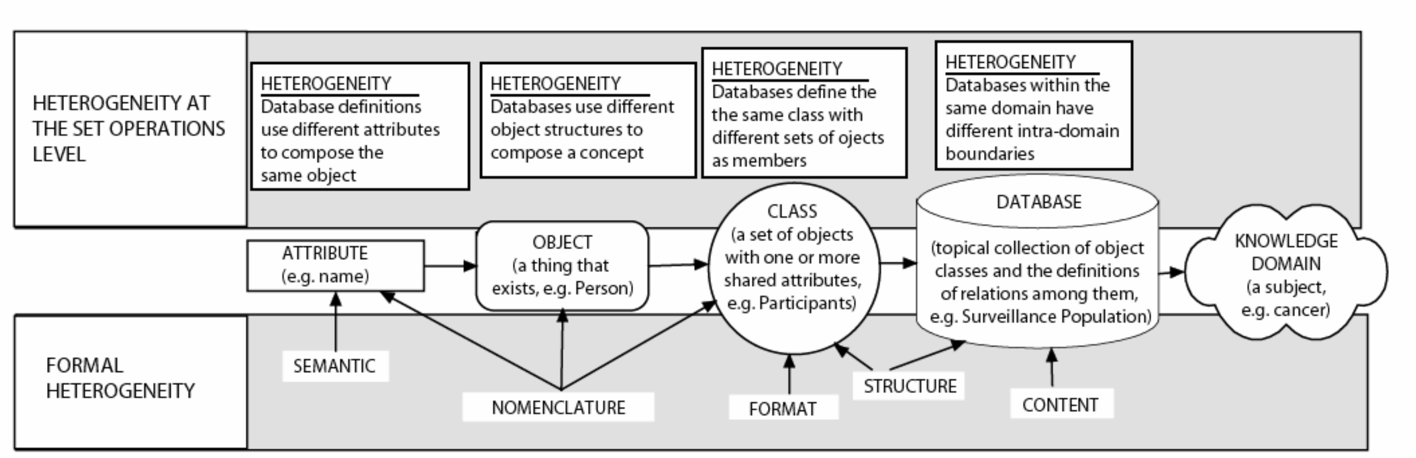
HETEROGENEITY AT THE SET OPERATIONS LEVEL
HETEROGENEITY Database definitions use different attributes to compose the same object
HETEROGENEITY Databases use different object structures to compose a concept
HETEROGENEITY Databases define the the same class with different sets of ojects as members
HETEROGENEITY Databases within the same domain have different intra-domain boundaries
ATTRIBUTE (e.g. name)
OBJECT (a thing that exists, e.g. Person)
CLASS (a set of objects with one or more shared attributes, e.g. Participants)
DATABASE (topical collection of object classes and the definitions of relations among them, e.g. Surveillance Population)
KNOWLEDGE DOMAIN (a subject, e.g. cancer)
FORMAL HETEROGENEITY
SEMANTIC
NOMENCLATURE
FORMAT
STRUCTURE
CONTENT

Figure 1. Data Heterogeneity Model

CGN Example

Some background about the nature of the collaborating entities helps to frame discussion of the details of data heterogeneity in a particular research network.

The CGN is made up of eight research centers, several of which are themselves collaboratives:

- Carolina-Georgia (Duke University, Emory University, University of North Carolina),
- Georgetown University, UCI-UCSD (University of California at Irvine and at San Diego),
- Mid-Atlantic (Johns Hopkins University and Greater Baltimore Medical Center),
- Northwest (Fred Hutchinson Cancer Center and University of Washington),
- Rocky Mountain (University of Utah, University of Colorado, University of New Mexico),
- Texas (M.D. Anderson Cancer Center, University of Texas Health Science Center, University of Texas Southwestern Medical Center, Baylor College of Medicine)
- University of Pennsylvania.

Each of these entities long pre-dates the establishment of the CGN research consortium, which implies that the design decisions that went into their informatics systems were not cognizant of the ultimate requirements of the CGN. For example, Duke University operates a comprehensive cancer center, while the University of California Irvine is a population-based cancer surveillance center. Consequently, with respect to data collection, Duke is oriented toward diagnosis and treatment information, while UCI is oriented toward population statistics. The data collection and data management processes at these two types of centers are different, as are the data confidentiality considerations. Creating a collaborative network with shared data from these two CRNs will necessitate bridging the gaps between their distinctly different research purposes.

Table 1 illustrates some potential sources of heterogeneity between a clinical center (e.g. Duke) and a population-based center (e.g. UCI). [Note: this table is for illustrative purposes only and does not represent the actual data structures of any particular institution.]

Source of Heterogeneity	**Clinical**	**Population-Based**
Set Operations		
Attribute Level	Identifier=Patient ID	Identifier=FamilyID+PersonID
Object Level	Case=PatientID + Physician ID + Diagnosis Code +Date Last Treated + Date Next Treatment	Case=PersonID + Source ID + Diagnosis Code + Vital Stat + Pathology Report
Class Level	Family=Patient + Spouse ID + Child 1..n	Family=All PersonID where FamilyID=FamilyID
Domain Level	Cancer data tables: Person, Physician, Diagnosis Code, Treatment Protocol, Tumor Type, Behavior	Cancer data tables: Family, Person, Tumor
Representational Forms		
Semantic	The attribute for patient last name='LNAME'	The attribute for person last name='Family_Name'
Nomenclature	Column Definition: 'Patient_ID' is the identifier for each discrete person who is affected by cancer.	Column Definition: 'Proband_ID' is the identifier for each discrete person who is affected by cancer.
Format	Patient ID is a string consisting of the first 6 characters of the patient last name plus the last four digits of the SSN.	Person ID is an integer consisting of the Family ID number plus the Person Number.
Structure	The database shall contain one table for ICD-O-3 disease classification codes and one table for SEER disease classification codes.	The database shall contain a diagnosis code cross-reference table that matches ICD-O-3 disease classification codes to SEER codes and recodes to a single synthetic disease classification code.
Content	The domain of inquiry is patients.	The domain of inquiry is tumors.

Table 1. Sources of Heterogeneity Between Two Network Nodes

As this example shows, there can be considerable differences between data resources that are part of the same general domain of inquiry, i.e. Cancer.

CGN Solution

The Cancer Genetics Network uses a star topology, or hub-and-spoke architecture, to connect its nodes. It should be noted, however, that the network nodes are physically connected to the hub only at the time of data transmission. Connections are not kept alive in the interim. At the informatics center (IC) core lies a standard relational database, which is the repository for all of the networked centers' data. Each network node maintains its own information systems and underlying databases. The content and structure of the nodes' databases are invisible to the IC. Data are collected from each node through an extraction, transmission, error resolution, re-transmission cycle. The network members have devised a set of common data elements through mutual agreement. These private CDEs have been codified in a manual available to all network members. The CDEs are organized by topic into table sets and validation rules, termed volumes. Some of these CDEs refer to published standards, such as the *International Classification of Diseases for Oncology,* Third Edition (ICD-O-3). However, most of the CDEs are specific to the Cancer Genetics Network. There is no general attempt to create a knowledge-based infrastructure for the data, for example, through use of an ontology.

The task of each CGN node is to extract the relevant data from its own repositories, then reformat and export it to the IC according to the agreed-upon standards. This is a non-trivial effort, due to the differences between data structures and standards prevailing at the node level vs. those mandated by the agreed network standards. For example, the clinical data repository at the Carolina-Georgia node contains over 400 tables from which data must be selected and re-formatted. These underlying data structures and their contents change over time. New patients are acquired and updates are applied to existing records. Therefore, each transmission from a network node to the IC is an extract of all relevant data, i.e. it is a cumulative export, not an incremental one. This distinction is important in its implications for error resolution. Each data transmission is subjected to extensive error checking routines upon arrival at the IC. These routines check the data for conformance to network data standards, however a more important consideration is data validation. Although each network node is presumed to be competent at manipulating its own data, it sometimes happens that changes to the data structure or content will introduce errors into the extracted product. Despite the fact that much of the data in each transmission is redundant to previous transmissions, each item must, therefore, be validated again upon receipt at the IC. The result of the error checking/data validation routines is an error report. A non-negotiable aspect of many clinical research networks is that locally collected data must be controlled locally. Data belongs to the research center that collected it, not to the network as a whole. Consequently, the informatics center has no authority to correct data errors in the transmissions from each node. In order to correct errors, the error report must be transmitted back to the node, where the individual research center is then responsible for finding and correcting the sources of the errors. Only when all errors are corrected is the data re-transmitted to the IC, whereupon it is again run through the error checking and validation routines where it is common to find additional errors, causing the cycle to be repeated. It is clear that this is a laborious and time-consuming process, exacerbated by the fact that each transmission is larger than the last.

INTEGRATIONAL DATA SHARING USING DATA STRUCTURE AND CONTENT DISCOVERY TOOLS

This paper proposes some means to overcome several of the inherent difficulties with the informatics tools used within existing clinical research networks, as exemplified by the CGN. Future clinical research networks will need careful attention to the informatics supporting the enterprise. Tools must be designed that facilitate, rather than inhibit, data sharing and knowledge production in these collaboratives. In particular, these tools must be able to provide answers to hypothesis-driven queries, but also, and more importantly, they must be able to support knowledge discovery activities. Discovering emergent knowledge held within a data repository requires the ability to reason about the contents. By contrast, hypothesis-driven queries require only the ability to retrieve information. Each has an important role to play in the knowledge production of a clinical research network and each has a different set of basic informatics requirements.

Following is an outline for a set of tools that meet the needs of both knowledge discovery and hypothesis-driven queries, and improve the informatics effectiveness of clinical research networks by taking a top-down approach to knowledge management and by eliminating the bottleneck inherent in a centralized, i.e. star network, informatics architecture. The term applied to the application of these proposed tools is the “integrational” approach, as differentiated from the centralized or distributed approaches to data management. It is termed integrational because there is no central database (ultimately) and because, although each network node will still collect and control its own data, other network members will share seamless access to the data and processing will be able to take place in a grid framework, i.e. where and when resources are available.

Semantic Analysis and Content Induction

The traditional relational model is designed to support well-defined queries, i.e. well-structured problems (Newell and Simon, 1972). Given a model of the problem, a query engine uses the internal structure of the database, its entity-relationship model, to compile a set of matching results. No reasoning is required on the part of the query agent, since it needs only its example and a map in order to produce its results. Such problems may be quite complex, but they are nevertheless highly structured. Overcoming data heterogeneity when solving such well-structured queries requires that the query engine be able to recognize data elements that represent the same object even though there are differences in the way that object is represented. As noted above, controlled vocabularies and common data elements can be used to mitigate data heterogeneity at the levels relevant to these well-defined queries.

Reaching an agreement on data element representation within a clinical research network is achievable, as has been shown by the adoption of network-specific “data dictionaries”. In addition, there exist controlled vocabularies in several knowledge domains of interest to NIH-sponsored clinical research networks, e.g. the Unified Medical Language System (UMLS) (Lindberg et al., 1993). Extending existing vocabularies into detailed knowledge domains, such as those that are specific to a given research network, can be done in several ways. Among these is the existing method of mutual agreement to a private definition. This, however, can be labor intensive and prolonged, is not designed for reuse, and can create conflict. Another way to generate the terms is through inductive processes, e.g. machine learning. In the integrational approach to clinical network informatics, common data elements are derived through cluster analysis of the data content of the participating nodes, as suggested by Dash, *et al* (Dash and Liu, 1997, Dash et al., 2001). This analysis may lead to a final data element definition or to an intermediate proposal for consideration by participants. If the latter, it at least has the face validity conferred by unbiased analysis.

The details of how the proposed cluster analysis is to be carried out are beyond the scope of this conceptual paper, but plans are underway for an experiment to test this method of vocabulary item induction.

Ontology Development and Structure Induction

Ontologies represent concepts and the relationships among them. Within the biomedical domain, ontologies "provide an organizational framework of the concepts involved…in a system of hierarchical and associative relations that allow reasoning" (Bodenreider et al., 2002). The formalization of these concepts and relationships helps to overcome inherent heterogeneity among data repositories in a collaborative by allowing the underlying contents of a given database to be mapped to a structure in the ontology and, through that mapping, to establish its relationship to other concepts. It also helps to overcome some of the fundamental limitations of the entity-relationship-attribute method of modeling knowledge of the world (Wand et al., 2000). One way to improve the efficiency of knowledge discovery within a clinical research network is through the use of knowledge discovery mediators (Wiederhold, 1992) or agents (Maamar and Moulin, 1997) that can carry out this mapping process independently. Creating and utilizing such agents is a valuable added efficiency to be extracted from a research network. As Farquhar, *et al* state, "ontology construction is difficult and time-consuming" and yet "essential for…the interoperation of heterogeneous systems" (Farquhar et al., 1995).

In the case of the clinical research network, there are existing ontologies and developing ontologies whose components should be reused where applicable. These include continuing work toward a general biomedical ontology based on the UMLS (Bodenreider, 2001), and a clinical trials ontology (Sim and Rennels, 1995). Where ontologies are not yet built out to meet the concepts required within a CNR's domain, it will be necessary to construct them. Kashyap and colleagues at the National Library of Medicine have described a partially automated method of inductive ontology construction (Kashyap et al., 2003). This taxonomic extraction method is expected to be tested and extended in the context of building a CRN informatics management toolkit.

CONCLUSION

This paper has examined the sources of data heterogeneity within clinical research networks (CNRs) and provided an overview of how, in an existing network, this problem is managed through reliance on a centralized informatics system and network-specific common data elements. The centralized approach is labor intensive and presents a single point of failure for the network. In addition, it inhibits knowledge production through the latency inherent in an iterative data processing cycle. The centralized approach currently in use was not designed to reuse existing components (which were not available at design time) and consequently it cannot easily capitalize on work being done in biomedical ontologies or the newer shared vocabularies. Finally, the centralized approach is somewhat inconsistent with the data sharing initiative recently announced by the National Institutes of Health. For these reasons, it is recommended that future clinical research networks' informatics systems be structured as an integrated ring topology where, theoretically, each member node has access to the network-common data set. There are different ways of designing such a structure, one of which is to create a network-specific mapping, such as was done for the Early Detection Research Network. Another is to take a fully integrational approach in which software agents, using an ontological framework to comprehend the structure and content of each contributory database, construct the mappings of the data elements using machine learning techniques, such as cluster analysis. The agents can and should run in either a continuous monitoring mode or on-demand so that changes to the member nodes' data structures are assimilated seamlessly. Part of the agents' functions would include notifications of changes to network members. These agents can be deployed in various ways depending on the physical network architecture in which they operate. Ideally, they would be hosted in a grid environment and be accessible to network members on demand and independent of the members' resident resources. While the latter approach is technologically daunting, current research efforts suggest that it is not impossible. To prove this concept, experiments are planned on two fronts: 1) data content analysis and inter-node semantic mapping, and 2) ontology extension through inductive methods.

ACKNOWLEDGMENTS

Parts of this work were supported by NCI grants 3U01CA078296-05S2, 5U01CA078296-05, and 5U01CA078285-05. The development of the EDRN data sharing tools was done by Dan Crichton and colleagues at the Jet Propulsion Laboratory, Pasadena, CA. I am grateful to Dan for sharing information on their efforts. Mark Peedin, Senior Analyst at the Duke Medical Center Cancer Center shared information on the structure of Duke's clinical database. The main database design and programming activities for the CFR and CGN Informatics Centers were done by Maureen Gildea-Schroeder, Informatics Manager, Epidemiology Division, University of California, Irvine. Hoda Anton-Culver, Ph.D., is Principal Investigator for the Informatics Centers of the CFR and the CGN Informatics Center at UCI. The NCI Program Manager for the CGN network is Carol Kasten-Sportes, MD.

REFERENCES

1. Bodenreider, O. (2001) Medical Ontology Research, in Lister Hill National Center for Biomedical Communications, Birmingham, AL.
2. Bodenreider, O., Mitchell, J. A. and McCray, A. (2002) Biomedical Ontologies, in *Pacific Symposium on Biocomputing*
3. CFR (1996) Cancer Family Registries: Overview, in *Cancer Family Registries Web Site*, Cancer Family Registries Informatics Center at University of California Irvine, http://www.cfr.epi.uci.edu.
4. Covitz, P. A., Hartel, F., Schaefer, C., De Coronado, S., Fragoso, G., Sahni, H., Gustafson, S. and Buetow, K. H. (2003) caCORE: A common infrastructure for cancer informatics, *Bioinformatics,* **19,** 2404-2412.
5. Crichton, D., Downing, G. J., Hughes, J. S., Kincaid, H. and Srivastava, S. (2001) An Interoperable Data Architecture for Data Exchange in a Biomedical Research Network, in *IEEE Computer-Based Medical Systems*, pp. 65-72.
6. Dash, M. and Liu, H. (1997) Similarity Detection among Data Files—A Machine Learning Approach, in *IEEE Knowledge and Data Engineering Exchange Workshop*
7. Dash, M., Tan, K. and Liu, H. (2001) Efficient Yet Accurate Clustering, in *IEEE International Conference on Data Mining*, Vol., pp. 99-106.
8. EDRN Early Detection Research Network, in *EDRN Web site*, 2003 National Cancer Institute, http://www3.cancer.gov/prevention/cbrg/edrn/organization.html.
9. Farquhar, A., Fikes, R., Pratt, W. and Rice, J. (1995) Collaborative Ontology Construction for Information Integration, Stanford University, Palo Alto.
10. Gilbert, W. (1991) Towards a paradigm shift in biology, *Nature,* **349,** 99.
11. Kashyap, V., Ramakrishnan, C. and Rindflesch, T. C. (2003) Towards (Semi-)automatic Generation of Bio-medical Ontologies, in *American Medical Informatics Association Symposium*, pp. 886.
12. Kincaid, H., Crichton, D., Winget, M., Kelly, S., Johnsey, D., Srivastava, S. and Thornquist, M. (2003) A National Virtual Specimen Database for Early Cancer Detection, in *IEEE Computer-Based Medical Systems*, pp. 117-123.
13. Lafky, D. (2003) Knowledge Integration for the post-genomic era: A progress report, in *IEEE Enterprise Networking and Computing in the Healthcare Industry (Healthcom)*, pp. 10-14.
14. Lindberg, D. A., Humphreys, B. L. and McCray, A. (1993) The unified medical language system, *Methods Inf. Med.,* **32,** 281-291.
15. Maamar, Z. and Moulin, B. (1997) An Agent-Based Approach for Intelligent and Cooperative Systems, in *IEEE Knowledge and Data Engineering Exchange Workshop*
16. Musen, M. A. (1992) Dimensions of knowledge sharing and reuse, *Computers and Biomedical Research,* **25,** 435-467.
17. Nass, S. J. and Stillman, B. W. (Eds.) (2003) *Large-Scale Biomedical Science,* Institute of Medicine, National Research Council of the National Academies, Washington, DC.
18. Newell, A. and Simon, H. (1972) *Human Problem Solving,* Prentice Hall, Englewood Cliffs, NJ.
19. Seminara, D. (1999) Innovative Study Designs and Analytic Approaches to the Genetic Epidemiology of Cancer, *J Natl Cancer Inst Monographs,* **26,** 0-1.
20. Sim, I. and Rennels, G. (1995) A Trial Bank Model for the Publication of Clinical Trials, in *19th Annual Symposium on Computer Applications in Medical Care*, New Orleans, LA, pp. 863-867.
21. Wand, Y., Storey, V. C. and Weber, R. (2000) An Ontological Analysis of the Relationship Construct in Conceptual Modeling, *ACM Transacations on Database Systems,* **24,** 494-528.
22. Wiederhold, G. (1992) Mediators in the architecture of future information systems, *IEEE Computer,* **25,** 38-49.
23. Zacks, R. (2001) Under biology's hood, *Technology Review,* **104,** 53-58.
24. Zerhouni, E. (2003) The NIH Roadmap, *Science,* **302,** 63-65.