Review Paper ■

Integration and Beyond:
Linking Information from Disparate Sources and into Workflow

William W. Stead, MD, Randolph A. Miller, MD, Mark A. Musen, MD, PhD, William R. Hersh, MD

Abstract The vision of integrating information—from a variety of sources, into the way people work, to improve decisions and process—is one of the cornerstones of biomedical informatics. Thoughts on how this vision might be realized have evolved as improvements in information and communication technologies, together with discoveries in biomedical informatics, and have changed the art of the possible. This review identified three distinct generations of "integration" projects. First-generation projects create a database and use it for multiple purposes. Second-generation projects integrate by bringing information from various sources together through enterprise information architecture. Third-generation projects inter-relate disparate but accessible information sources to provide the appearance of integration. The review suggests that the ideas developed in the earlier generations have not been supplanted by ideas from subsequent generations. Instead, the ideas represent a continuum of progress along the three dimensions of workflow, structure, and extraction.

■ **JAMIA.** 2000;7:135–145.

The mission of biomedical informatics is to enable people to use information to improve health. This mission, coupled with visions of how information might improve health care process, health outcomes, and biomedical research, has not varied greatly since the inception of the field. The vision of integrating information from a variety of sources into the way people work, to improve decisions and process, has been one beacon.[1] Although the goal has not changed, our understanding of how it might be achieved has evolved. Improvements in information and communication technologies, together with discoveries in biomedical informatics, have changed the art of the possible.

This paper traces the evolution in thinking by drawing a distinction among three generations of work. First-generation projects build everything required by the application environment de novo. Integration in first-generation systems is a natural by-product of the use of a self-contained system with a single database for as many purposes as possible. Second-generation projects integrate information from various sources and systems. Second-generation integration brings information together through enterprise information architecture. Third-generation projects inter-relate disparate but accessible information sources using techniques ranging from information structures to extraction. Third-generation projects provide the ap-

Affiliations of the authors: Vanderbilt University, Nashville, Tennessee (WWS, RAM); Stanford University Medical School, Stanford, California (MAM); Oregon Health Sciences University, Portland, Oregon (WRH)

This paper was presented in part as the keynote to the Cornerstone on Integrating Information, one of four Cornerstone sessions included in the program of the AMIA Annual Fall Symposium, Washington, DC, November 6–10, 1999.

The work leading to the conceptualization of the three generations and three dimensions was supported in part by grants G08-LM04613 and G08-LM05443 from the National Library of Medicine (NLM); the work on provider order entry, by grants R01-LM06226 and G08-LM05443 from the NLM; the work on component-based architectures for decision support, by grant LM05708 from the NLM; the work on Cliniweb, by grant 1U01-LM05879 from the NLM and grant DE-FG03-94ER61918 from the Department of Energy; and the metadata work, by training grant LM07088 from the NLM.

Correspondence and reprints: William W. Stead, MD, Associate Vice Chancellor for Health Affairs, Vanderbilt University, 416 Eskind Biomedical Library, 2209 Garland Avenue, Nashville, TN 37232-8340; e-mail: ⟨bill.stead@mcmail.vanderbilt.edu⟩.

Received for publication: 11/1/99; accepted for publication: 11/18/99.

pearance of integration while preserving content richness through diversity.

Over the course of this evolution, some ideas have stood the test of time. Others have proved to be unworkable and have limited progress. In time, new ideas have appeared and, in combination with proven concepts, have pushed back the barriers. The cycle has then started over as new limits are reached. On reflection, the proven concepts represent a continuum of progress along the three dimensions of workflow, structure, and extraction. This paper concludes with comments about the state of the art in each of these dimensions.

Evolution: From Creating, through Integrating, to Relating

The first generation began in the late 1960s. Pioneers had already shown that it was possible to represent clinical information in a digital computer.[2] However, those ''proof of concept'' projects did not turn into working, clinical applications. The technology of the day did not support data capture and use during the health care process. The minicomputer[3] broke this barrier and led to first-generation ''integration'' projects.

First-generation projects had to *create* databases. They had to capture the information they needed and then use it to do something that would otherwise be done by a person. The automated history taker, developed at Duke University in 1970, is an example of an early first-generation project. An interactive, adaptive questioning program obtained the history of a patient complaining of headache, and used rules to make a diagnosis.[4] Benefits included a detailed and legible history, reduction in the physician time needed to obtain the history, and an expert opinion on par with that of a neurologist. Although the benefits were real, they were of narrow scope and the costs were such that use in routine operation was not justified.

The obstetrical medical record system[5] developed at Duke University is an example of a mid–first-generation project. A variety of automated history and physical examination takers were combined to create a prenatal record. The data from the initial workup and followup visits could be re-used to simplify admission notes and discharge summaries throughout the course of the pregnancy. The benefits of this reuse were enough to justify use of the application in an operational mode at its development site for 30 years. However, the application used program code to reproduce the paper documents that had preceded it. As a result, it did not generalize into a computer-based patient record or transfer to other sites.

The Medical Record (TMR)[6] is an example of a mature first-generation project. Prior to implementation, a health facility's data requirements are modeled and reflected in a dictionary of rich metadata. Patient data are captured at the source, refined, and augmented throughout the process of providing patient care. For example, data captured in appointment scheduling provides shortcuts during registration and check-in. Similarly, clinical data captured in sufficient detail to support report generation for the procedure (such as a cardiac catheterization) provide more complete charge capture than those captured for a billing system. Direct cost of the system is offset by practice management efficiencies. A computer-based patient record is created as a by-product of practice management. This record supports direct improvements in the care of individual patients and indirect improvements as a tool for outcome and health services research.

The TMR project, and others like it across the country,[7–9] tested ideas that have stood the further test of time. Data can be captured at the source and refined through re-use. Role-specific displays can target presentation to individual need. Granular data can be used for any number of purposes if they are structured according to meaning. Metadata tables can be used to separate management of data meaning from application code. Attribute–value structures can optimize the management of sparse data.

Despite these successes, widespread implementation did not follow. Several key barriers were apparent as this generation began to phase out. The effort and expertise required to populate facility-specific metadata dictionaries were too great. The precision with which content meaning could be mapped between various systems was inadequate. The changes in workflow that were required to capture structured data throughout the care process were too great.

In short, first-generation projects provided integration by using a single system for all functions. No single system could optimally support all user roles. No single system could be populated by all the needed data or information. A pioneer spirit was therefore required to demonstrate the benefits of integrating data from a variety of sources into workflow in this fashion.

Second-generation projects began to appear in the 1980s, enabled by local area network technology. These projects *integrate* data and information across various systems to overcome the barriers identified through first-generation projects. StatLan[10] is an early second-generation project. Instead of a monolithic

hospital information system, each department, such as the laboratory or the pharmacy, installs a ''best of breed'' system for their area. Registration and admitting, discharge, transfer transactions are captured centrally and passed to participating systems. A patient view provides access to the information about a patient in each participating system. The ''IAIMS menu,''[11] piloted in sites implementing the first wave of Integrated Academic Information Management Systems, is another example. A single user interface provides access to personal productivity tools, MEDLINE, clinical systems, and other resources.

The early second-generation systems increased usability by providing access to the collective set of functions of the interconnected systems. However, each system was still a discrete unit and was managed as such. It retained its own data model, business logic, and database. The increase in functionality came at the price of redundant implementation effort, inconsistent definition of content across the systems being used, and inability to resolve discrepancies. In other words, the collection of databases did not equal an integrated database in the early second-generation models.

Data interchange standards such as Health Level Seven (HL7)[12] were the first step toward decreasing the effort required to integrate management across separate systems. The HL7 standard defines the interchange format, messages, and triggers. An application programmed to support this standard can interchange messages with other applications that observe the standard, without dedicated programming. Differences in data content between applications still require significant mapping. In rare cases, worldwide use of a standard data set avoids this problem. The Visual Human dataset[13] is an example of the concept. All derivative works based on this standard data set are interchangeable. The Logical Observations Identifiers, Names, and Codes (LOINC) codes[14] have the potential to become another example, if system implementers use LOINC codes as the actual codes for observations. If, instead, they associate the LOINC code for intersystem mapping with a local code for storage, it would be considered a third-generation strategy because of the potential ambiguity.

More mature second-generation projects sought to overcome this problem through an enterprise information architecture.[15] The MCIS-1 architecture[16] is an example. It manages the data model, metadata, knowledge bases, and databases as enterprise information resources. These information resources are managed separately from the applications that process transactions in the various facilities or departments that make up the enterprise. Communication management engines handle data distribution, request brokerage, serialization, and logical unit of work for applications. In this architecture, transaction-processing applications are managed as components supported by the common foundation of enterprise information resources. Role-specific user interfaces provide context-sensitive views of the information, capture data and decisions from the user, and hand transactions to the appropriate transaction-processing engine for action.

Collectively, the user interfaces, information resources, communication management processes, and application components make up an enterprise information system. Rich function is provided over a foundation of integrated data. The alignment of data across systems that derives from the use of common metadata, and the integrated databases that provide a single source to information, are the key differences between early and late second-generation projects. Numerous application-to-application interfaces are replaced by a single connection between an application and the communication management engine. Legacy applications are encapsulated by populating profiles with content from the enterprise information resources and managing their databases as temporary cache.[17] Components can be swapped, as new ones become available without restarting data modeling and metadata definition from scratch.

A number of academic sites and commercial products use these enterprise information architecture concepts to varying degrees.[18–20] At this stage in the second generation, a number of barriers again limit progress. Few commercially available products are designed to interoperate in this fashion, and expertise is required on site to put the pieces together. The nature and scope of most health enterprises are in constant flux. Quality improvement rests on measures from each facility a patient uses. The most powerful use of information technology is to make possible new relationships between customers and the enterprise. The tincture of time might overcome the first of these barriers, but the latter three suggest fundamental change in emphasis. Put simply, integration within the enterprise is necessary, but it is not likely to be sufficient.

Since the late 1980s, informatics research and development projects have provided building blocks for a third generation of projects. These projects explicitly *relate* otherwise separate data and information resources. In other words, data and knowledge that are outside a system or enterprise may be linked to the data and work processes that are within it. The Unified Medical Language System (UMLS)[21] is an example of a third-generation project. The UMLS Metathe-

Table 1 ■

Mapping of Proven Concepts from Three Generations along Three Dimensions

Dimension	Generation 1	Generation 2	Generation 3
Workflow	Source data capture Role-specific displays Decision support Engage customer directly		Distributed knowledge Distributed work process
Structure	Multi-use data models Metadata tables Attribute–value data structures	Interchange standards Standard data Information architecture Componentized software	Mark-up languages Object request broker architecture
Extraction			Relationship among data Approximation of structured data Machine learning Data filters

saurus contains the terms from source vocabularies, together with an explicit many-to-many mapping between terms. The set of relationships can be mined to provide multifaceted definitions. Similarly, the frequency with which two drugs occur in the same article coded in MeSH as ''adverse effects'' can, for example, be mined to infer the likelihood of a drug–drug interaction. Markup languages such as SGML[22] and XML[23] represent a different strategy. Tags such as the uniform resource identifier and the document type definitions provide a recognizable way to identify content. Object request broker architectures are a third approach. Objects can recognize each other and interoperate since everything but a specified public interface is internalized.

These projects add value by achieving linkage without the difficulty of direct integration. They make possible discovery through retention of diversity. In addition, the documented relationships between sources provide information that is not in any of the sources. In other words, the whole is greater than the sum of the parts. Global use is practical.

These advantages come at a price. Undisciplined diversity adds cost. Translation between separate sources will always be ambiguous. The decoupling of form and content limits quality control.

Transformation: From Generations to Dimensions

This review identifies ideas and techniques from all three generations that have stood the test of time. Each has advantages and limits. All coexist in successful modern projects. Table 1 shows the proven concepts from across the generations as a continuum of progress along three dimensions.

The first dimension involves workflow—data capture, communication, visualization, decision support, role modification, and change facilitation. The second dimension is about structures—to represent data without ambiguity and support regularization of content and componentization of software. The third dimension involves extraction—exploring data to discover information or knowledge.

Each situation requires a different balance of ideas from the dimensions. For example, the techniques involving structure can be exploited wherever homogeneity is practical. Extraction is needed when crossing boundaries. An ideal scenario involves use of extraction to implement structure, obtaining the benefits of homogeneity while decreasing what has to be done inside the enterprise and preparing for a change in boundaries. Similarly, use of structure and extraction reduce the need for data capture, increasing the amount of effort that can be devoted to the residual data capture requirement.

Figure 1 indicates when data should be captured in a structured form—when the data can be represented in a general-purpose data model, recorded at a granular level, according to standard metadata, and captured in a work process where the reward exceeds effort. Otherwise, data can be captured as computer-readable text or images and archived for subsequent information extraction.

Extraction techniques can be used to derive additional information from relationships among structured data and to make an approximation of the structured data

contained in unstructured text and images. Data filtering can focus views and query results to context.

Data can be captured at the source if privacy is preserved and if the user perceives a reward greater than the effort. The linkage of information use to data capture during the decision process can help justify the data capture effort. A change in roles, such as empowering customers to track their own progress, can shift the reward-to-effort balance in the right direction. If a positive reward-to-effort balance can not be obtained, consider leaving the data out of the computer-based record. If the data are essential, capture them in the least invasive manner possible. Before long, advances in information or communication technology will change the art of the possible.

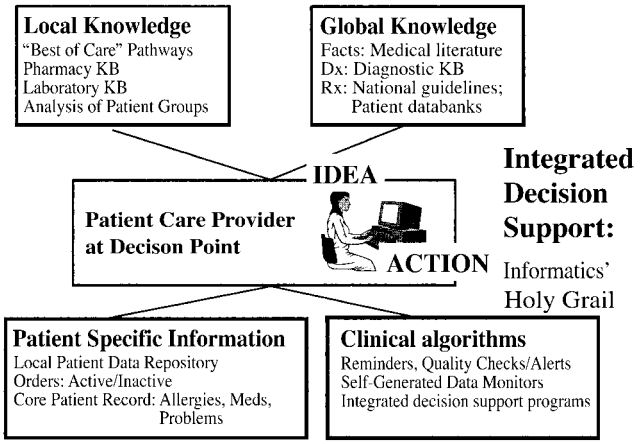

Figure 2 Integration of information into clinical workflow. KB indicates knowledge base.

Information Workflow Integration: Provider Order Entry as an Example

The critical factor in implementing clinical decision support systems is to do so in a manner that makes a measurable impact on patient care processes and outcomes. Figure 2 illustrates how this can be accomplished—by intervening at the time that care providers convert their ideas into actions. Order writing is the optimal time to bring additional information to bear on decision making. It must be recognized, however, that of the work required to implement significant clinical systems, such as care provider order entry, at least 75 percent is in social engineering and only 25 percent is technical. The organization and delivery of information in the workplace can facilitate the "social engineering" as well as the "technical" aspects of decision support.

Examples from the Vanderbilt University Medical Center environment illustrate the principles shown in Figure 2. Vanderbilt's philosophy is that any information required for workflow integration should, ideally, be represented outside all dedicated individual systems (e.g., in an institutional database that is not part of any legacy system). The concept has been given the name "Vanderbilt Externalized General Extensions Table," or VEGETABLE, locally. By representing business rules and other information externally to any dedicated system, this collective content can be made accessible to all users, and the chaos created by version changes and vendor changes in dedicated systems can be diminished. Data to support the following processes should, ideally, be externalized—identification, authentication, authorization, settings, roles, distributed capture of clinical expertise, links to the clinical data repository, event notification, links to knowledge resources, system troubleshooting (e.g., user log files), and system evolution (change history).

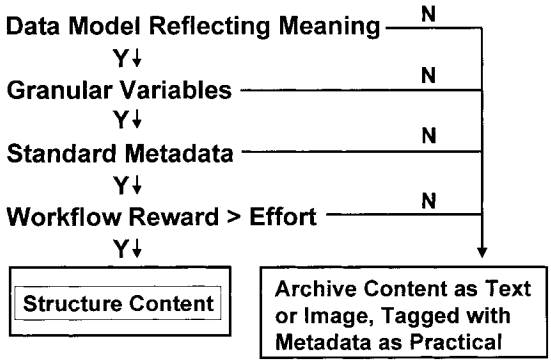

Figure 1 Decision tree for choosing between capture of structured, coded data and capture of data as text or image.

In Figure 2, as the practitioner sits at the point of care, a number of local knowledge resources can be brought to bear. First, many institutions have constructed local "best of care" clinical pathways for a number of specific clinical problems. To do so, institutions use their own expert doctors, nurses, and ancillary care providers, knowledge of local strengths and weaknesses, and information gleaned from both national guidelines and local databases. The ability to activate (on an item-by-item basis) such protocols ("order sets") within an order entry system can help provide more uniform, cost-effective care of patients. For example, a leading Vanderbilt physician on the bone marrow transplant unit developed a number of order sets—e.g., several protocols for the treatment of graft-versus-host disease—for the WizOrder clinician order entry program,[17] to standardize and improve the care of patients on that unit. These protocols have the order entry program calculate ideal body weight

surface area, adjusted ideal body weight, and estimated creatinine clearance as parameters for determining accurate doses for chemotherapeutic agents, based on milligram-per-kilogram or milligram-per-square-meter dosing guidelines. The protocols also include automatic ordering of critical items in association with other items (such as leucovorin rescue for high-dose methotrexate, or drug levels three times per week when cyclosporine is given), so that they cannot be omitted due to human error.

Another kind of local information is represented in hospital policies and procedures. Web pages in the WizOrder system allow users to both review electronic versions of Vanderbilt policies and procedures and activate orders related to specific procedures or guidelines through Web-based forms. A specialized instance of local procedural ordering occurs in the "healthy" newborn infant nursery, where nurses can activate an individual pediatric attending physician's "standing orders" for the care of a newborn infant admitted to their service (once the newborn infant has been examined and declared healthy by the obstetrician and the pediatric house staff) from a Web-based local protocol form (which can also be reviewed by the physician from the Web).

"Global" information can take the form of regional, national, or international databases or Web-based textural or multimedia resources. In the WizOrder system at Vanderbilt, entering orders for "diagnosis" and medication orders triggers UMLS-based mapping to clinical concepts (and to MeSH), so that literature about treating patients with the given diagnoses using medications similar or identical to the patient's medications can be available at "the click of a mouse." Similarly, as part of the National Library of Medicine–funded PC-POETS project at Vanderbilt, linkages to a medical diagnostic knowledge base and decision support system have been created.

Use of "patient-specific" information (Figure 2) can facilitate workflow integration in a number of ways. Through linkages from the patient-care-provider order entry system to the clinical data repository, it is possible to generate "lab alerts" whenever a user reviews the orders for a given patient (or, alternatively, to deliver lab alerts in real time through a notification engine). Not only are clinically significant "lab value out of bounds" alerts generated but, as part of the PC-POETS project, alerts are also generated when a trend in recent laboratory results suggests that the laboratory test values will be "out of bounds" within the next 72 hours. Additional alerts are generated when a test with results in the "normal" range (i.e., with a trend to remain "normal") has been ordered too many times. An additional use of the electronic availability of currently active orders and recent laboratory and textual report results is to be able to generate, for a given user's patient census list, a "current medications and results" (CMR) report that is limited to cover two sides of a single sheet of paper (per patient). The CMR reports make it easier, during ward team rounding sessions, for everyone to be "on the same page." More time can be devoted to discussing patient care and to teaching, since less time is devoted to reporting the laboratory and x-ray results that are summarized on the CMR printouts.

Finally, it is possible to invoke clinical algorithms in a patient-specific manner to provide decision support capabilities that facilitate workflow integration. By externalizing the pharmacy databases for drug interactions, drug-class based allergy alerts, and pharmacy monographs, it is possible to alert physicians to a patient's drug allergies and significant drug interactions instantaneously as medication orders are entered into the order entry system. Not only do such alerts prevent adverse drug events (ADEs) and decrease morbidity and costs, they also reduce the burden on the pharmacists of having to notify physicians of such ADE alerts when they are generated "downstream" by the pharmacy system, since physicians have already reacted to the alerts at the level of the order entry system. Clinical algorithms can also automatically determine patient eligibility for guideline-based care and provide convenient Web-based forms to allow physicians to order related protocol-based medications, tests, and nursing care.

Introducing clinical systems into the clinical arena involves a substantial degree of social engineering. By leveraging information resources of global, local, and patient-specific natures and by integrating useful clinical algorithms into the environment, the problems of workflow integration can be made easier. A system for distributed knowledge maintenance facilitates local experts who maintain portions of the overall system knowledge on the basis of their individual expertise.[24] By making decision support available when ideas are converted into actions (e.g., during clinician order entry), such a system can have a major impact on clinical processes and outcomes.

Ontologies: Data Representations to Support Linkages

As architectures for clinical information systems have become more distributed, workers in medical informatics have paid increasing attention to the problem of ensuring that all software components share a com-

mon data model. In the 1980s, developers began to grapple with the problem of providing systems for patient registration, for hospital pharmacies, and for clinical laboratories, among others, with a uniform view of the data that they processed.[25] In the 1990s, as software engineers began to build systems from components of increasingly fine grain size,[26] the problems of data integration became even more acute. In all cases, developers have needed to ensure linkages among different components by building explicit, inspectable, and editable representations of the data that the individual modules must process and communicate to one another.

The distributed software architectures that became popular in the 1980s required each departmental system to communicate with every other system on the network using a common data model. Cimino's work at Columbia–Presbyterian Medical Center (CPMC) to develop the Medical Entities Dictionary (MED) provides one of the best demonstrations of how a central terminology resource can facilitate interoperation among diverse subsystems.[25] The MED provides a lingua franca for expressing concepts related to patient care at CPMC and serves as the basis by which specific terminologic strings or codes used in legacy information systems can be translated into a format that other departmental systems can process. The MED was constructed pragmatically, on the basis of the particular data requirements of the specific systems in use at CPMC. The success of the approach is thus measured in terms of the tremendous degree of interoperation that the different systems in use at CPMC have achieved.

The MED represents an ontology, a predefined set of concepts, relationships among concepts, and constraints on those concepts.[27,28] The term *ontology* has been co-opted by workers in computer science from the branch of metaphysics that concerns the nature of existence; when software engineers speak of an ontology, they are referring to an explicit representation of the concepts that system builders define to exist in a particular domain. Just as the MED enumerates all the concepts that are relevant for data interchange at CPMC, ontologies in general provide reference models that specify all the concepts that may constitute the domain of discourse for specific information systems. As we think about the linkages required to support data exchange throughout the clinical enterprise, the notion of ontologies becomes central. Ontologies are no longer just arcane theories proposed by philosophers or obtuse components of computer programs built by workers in artificial intelligence; rather, ontologies form the basis for human–computer interaction and for information exchange throughout the health care environment. It is, therefore, not surprising that the notion of ontologies is receiving so much attention in the medical informatics community.

Ontologies Are for People

Ontologies allow software developers to conceptualize and formalize what they know about an application domain. Such data structures provide more information than do the paper-based data models that invariably result from the software engineering process—or even than do the online models produced by many tools for computer-assisted software engineering. Ontologies not only are understandable by both humans and computers but also typically document constraints among concepts and assumptions in the model in a manner that goes well beyond that associated with less formal approaches. For software developers, ontologies thus represent resources for describing everything about the application area that is relevant for the software engineering process.[29] They provide a human-readable and machine-processable description of the data modeling assumptions that all computer systems in the enterprise must address if they are to interoperate.

Most lay people have their most direct interaction with formal ontologies as a consequence of searching the World Wide Web. When users search for topics on the Web by browsing through fixed categories of information, they do so by browsing through an ontology. At Yahoo!, for example, the fourth employee to be hired by the company was designated the "chief ontologist," in recognition of the central role of a large, explicit ontology in driving the Yahoo! resource. All Internet resources that categorize Web pages do so by defining a corresponding ontology of topics.

Ontologies have become a focal point in much Internet-based commerce. The success of many online vendors such as Amazon.com has been attributed directly to the usefulness of their taxonomies of products. A rich ontology that makes fine distinctions among products allows online shoppers to locate the specific merchandise that they are seeking. On the other hand, a Web site that uses ambiguous or unfamiliar terms to categorize its offerings may inhibit potential customers from finding what they want. A remarkable feature of the Internet age is that, suddenly, the commercial success of many ventures is determined directly by the value of the ontologies that allow consumers to be linked to the specific products in which they are interested. Ontologies have moved from the realm of philosophers to the realm of entrepreneurs.

Although the term *ontology* has become a new buzzword, ontologies have always played a key role in human–computer interaction. Ontologies, either explicitly or implicitly, define the domain of discourse for linking computers and people.[30] Whenever users interact with any computer program, they must share with the program's developers an ontology of the terms presented in the user interface. Without a shared ontology, it is impossible for users to interpret the computer's behavior and to know what terms to enter or what menu selections to make to achieve their goals. When users communicate with a program such as Quick Medical Reference (QMR),[31] for example, they do so only by internalizing QMR's ontology of diseases and their manifestations and by elucidating the computer's actions in terms of that ontology.[27] When system builders make ontologies explicit, they make it easier for users to interact with the computer. The more system builders can formalize the semantics of the concepts needed for interaction with their programs, the better assurance users can have that they will communicate successfully with the computer. The more formal the ontology, the more the developers also can prevent the meaning of concepts from drifting over time.

Ontologies Are for Computers

As we move to third-generation distributed systems, ontologies play an increasingly important role in linking computer programs with other computer programs. Just as the MED[25] allows the different departmental systems at CPMC to interchange information through a canonic set of concept representations, ontologies permit modern component-based software systems to refer to a single, sanctioned description of the types of data on which they operate.[26] As a result, all the components share a data model that they can obtain dynamically from a single server. Thus, generic software components can acquire their domain-specific functionality by interacting with a relevant ontology that defines the salient concepts in the application area. New generic components can be ''plugged into'' the architecture and can obtain all necessary domain descriptions from the shared ontology. Most important, domain-specific concepts are defined in only one place, and all changes to the ontology that developers make over time are automatically made available to the component software modules without the need for any reprogramming.

A good example of this kind of third-generation architecture is EON, developed at Stanford University to provide decision support for protocol-based medical care.[32] EON contains a software module that determines appropriate therapy, given a patient condition and a protocol according to which the patient should be treated, and another module that identifies new protocols or guidelines for which a particular patient might be eligible. Both software components obtain all their domain-specific knowledge of the relevant medical specialty from an explicit ontology. When that ontology defines the kind of laboratory tests and clinical interventions that are appropriate for, say, therapy of HIV-related disease, then these two software components can recommend to the clinician indicated HIV medications and can identify appropriate new protocols for antiretroviral therapy or for management of opportunistic infections. When the ontology instead includes the laboratory tests and clinical interventions most appropriate for, say, breast cancer, then the software modules perform their tasks in a manner suitable for this other domain. The ontology is easy for developers to edit, and extension of the EON architecture to new medical specialties becomes a matter of modifying the ontology to define the concepts that are required for the new application domains.

Just as ontologies play a key role in human–computer interaction, they are taking center stage in supporting the *computer–computer* interactions that are needed to enable a variety of Internet-based applications. Particularly in electronic commerce, where computers may interact with one another autonomously to identify online vendors who can provide necessary products and services, and where purchases and payments may take place without direct human intervention, there must be an explicit means to mediate the dialog between different Internet-based software agents. To enable such interoperability, a working group supported by the companies who participate in CommerceNet recently released the eCo framework for electronic commerce.[33] The eCo framework provides a single common protocol through which software agents can describe their features, services, and interoperability requirements. The means by which software agents perform these tasks is by exchanging specific ontologies defined in XML.

The needs of electronic commerce over the Internet at first may seem far removed from the data-processing requirements of health care institutions. Nevertheless, health care organizations are inexorably becoming more complex and more distributed. With the rapid advent of Internet-based systems for applications such as consumer health, claims processing, telemedicine, and regional data integration, there is already an accelerating need for third-generation systems that can exchange information in flexible and adaptable ways. Proposals for data linkages for electronic commerce are highly relevant to workers in medical informatics.

The ability of software agents to negotiate among themselves and to deliver services to one another will become central to the functioning of clinical information systems as they reach out to remote care sites, to patients' homes, and to a wide variety of payers and vendors. Explicit ontologies that allow software components to communicate their functionality and to declare what those components assume about the world will be essential elements of those third-generation systems.

Extraction: Mining and Filtering

The integration of medical information can be achieved in part by information extraction through data mining and filtering. To understand the integration of information using data mining and filtering, we must first understand the two types of information in health care and operations that apply to each. There are two types of information in health care: patient-specific and knowledge-based. Patient-specific information is information generated in the care of patients, i.e., the medical record and its associated text, images, and codes. This is distinguished from knowledge-based information, which is the scientific literature of health care and its derivative resources.

The value of patient-specific information is in the aggregation of individual and group data. The procedure to extract information is *data mining*. The benefit of knowledge-based information is the critical application of pertinent and best-quality literature. The procedure to do this is *data filtering*.

Data Mining

Data mining is "the use of historical data to discover regularities and improve future decisions."[34] It applies techniques from machine learning, Bayesian statistics, and such. It is also called "knowledge discovery from databases" (KDD), although some have warned that its uncritical application might also lead "data dredging."

How can data mining provide value in health care? Its most important value may be in the discovery of *ir*regularities, such as variations in clinical practices across regions and practitioners,[35] profiling of practitioners and health plans, and the discovery of excess costs for diagnostic testing, pharmaceuticals, and so forth. Data mining can also lead to the discovery of patterns to guide research. For example, it can discover associations among diseases and their etiologies. It may make possible the realization of a project proposed in the 1980s to discover hypotheses from clinical databases in an automated fashion to guide research, the RADIX Project.[36]

What are the impediments to and limitations of data mining? Some may question its retrospective analysis, which can lead to undetected biases. Another limitation is that in claims databases, the clinical medical record, and other operational systems, some elements of the data may be missing, incomplete, or otherwise inadequate.[37] Data mining is also limited by two factors that impede the use of data in many medical informatics applications. First, it is hampered by current clinical vocabularies that limit the ability to express medical concepts.[38] Furthermore, many data are "locked" in narrative text, with little hope of extraction.[39]

Data Filtering

How can we find the best information to apply in medical decision making? We must improve the production of and access both to the primary scientific literature, i.e., the original research reports typically published in medical journals, and to the derived literature, including review articles, textbooks, practice guidelines, and other syntheses of scientific knowledge.

The access to knowledge-based information is limited by current approaches for several reasons. First, the primary research on any topic is scattered about the literature. This is often done deliberately, as authors try to spread their publications across a diverse number of journals.[40] The problem is that it makes retrieval for synthesis more difficult. Another problem is that metadata structures are inadequate to retrieve all the potential information on a topic, as evidenced by studies showing that few MEDLINE search strategies allow complete retrieval of articles for general searches[41] or systematic reviews.[42] In addition, few physicians or others are skilled in appraising the evidence for clinical decisions. We therefore need better means to produce and access both primary and derived literature.

Improving the production of primary literature should center on improving access to data. The model for doing this should be inspired by the bioinformatics community, which has a history of data sharing among investigators.[43] Sharing data may, of course, be more problematic, because of the potential for violation of personal privacy. Appropriate de-identification of information should be possible, however.[44] The recent efforts by the National Institutes of Health (NIH) to create standardization process for clinical trials data collection should help, as should the electronic publishing of supplementary data.

Access to primary literature is facilitated by better organization and indexing of content. The beginnings of

this can be achieved by improved development and use of metadata to facilitate retrieval. This will probably be helped by automated indexing approaches that make such indexing easier.

Improving the production of derived literature is important for applying scientific knowledge to evidence-based care. The existing derived literature, however, is neither as comprehensive nor as evidence-based as we might hope.[45] The Internet in particular gives rise to low-quality information.[46,47] Anyone can be a publisher, which is good for a democratic society but potentially problematic in professions such as medicine. We need to see more evidence-based content for practitioners and consumers as well as a better understanding of health and research processes.

To improve access to the derived literature, we must improve access to quality information. This can be done by a variety of means:

- Voluntary codes of conduct, e.g., Health on the Net (HON, at www.hon.ch) guidelines
- Application of established criteria, e.g., the elements specified by Silberg et al.[46]
- Catalogs that filter for high-quality information, e.g., CliniWeb[48]
- Algorithmic approaches to determining quality content, e.g., those described by Price and Hersh[49]

To integrate information, we must also integrate data mining and filtering. The filtering of patient-specific data may allow more effective mining. Mining of filtered knowledge-based information may show us new directions for future research.

Conclusions

To integrate is to bring parts together into a whole. *Integrate* is synonymous with ''combine,'' ''equalize,'' ''blend,'' and ''regularize.'' To relate is to bring into or establish association. *Relate* is synonymous with ''link,'' ''get in touch with,'' and ''compare.'' Integration achieves clarity at the cost of the potential for discovery through diversity. Recent work has shifted from a focus on integration alone to a balance between integration and relation. Over time, research may enable the best of both worlds. Standard data may permit representation of diversity while providing the substratum for integration. The intersection of the results from a suite of complimentary extraction techniques applied to various related data may provide answers that have the clarity of the results of a query to an integrated information resource.

References ■

1. Stead WW. Information Technology Video: Scenarios 2 [videotape]. Nashville, Tenn: Vanderbilt University Medical Center, 1998.
2. Stead WW. A quarter century of computer-based medical records. MD Comput. 1989;6(2):74–81.
3. Stead EA. Stead WW. Computers and medical practice: old dreams and current realities. MD Comput. 1985;2(6):26–31.
4. Stead WW, Heyman A, Thompson HK, Hammond WE. Computer-assisted interview of patients with functional headache. Arch Intern Med. 1972;129:950–5.
5. Stead WW, Brame RG, Hammond WE, Jelovsek FR, Estes EH Jr, Parker RT. A computerized obstetrical record. Obstet Gynecol. 1977;49:502–9.
6. Stead WW, Hammond WE. Computer-based medical records: the centerpiece of TMR. MD Comput. 1988;5(5):48–62.
7. McDonald CJ, Blevins L, Tierney WM, Martin DK. The Regenstrief Medical Records. MD Comput. 1988;5:34–47.
8. Pryor TA. The HELP medical record system. MD Comput. 1988;5:22–3.
9. Barnett GO. The application of computer-based medical record systems in ambulatory practice. N Engl J Med. 1984; 310:1643–50.
10. Simborg DW, Chadwick M, Whiting-O'Keefe QE, Tolchin SG, Kahn SA, Bergan ES. Local area networks and the hospital. Comput Biomed Res. 1983;16:247–59.
11. Sengupta S. Heterogeneity in health care computing environments. Proc 13th Annu Symp Comput Appl Med Care. 1989:355–9.
12. Health Level Seven Vocabulary, Standard Version 2.3: An Application Protocol for Electronic Data Exchange in Healthcare Environments. Ann Arbor, Mich: Health Level Seven, 1994.
13. Spitzer V, Ackerman MJ, Scherzinger AL, Whitlock D. The Visible Human Male: a technical report. J Am Med Inform Assoc. 1996;3:118–30.
14. Huff SM, Rocha RA. McDonald CJ, et al. Development of the Logical Observations Identifiers, Names, and Codes (LOINC) vocabulary. J Am Med Inform Assoc. 1998;5:276–92.
15. Hripcsak G. IAIMS architecture. J Am Med Inform Assoc. 1997;4(2 suppl):S20–30.
16. Stead WW, Borden RB, Boyarsky MW, et al. A system's architecture which dissociates management of shared data and end-user function. Proc 15th Symp Comput Appl Med Care. 1991:475–80.
17. Geissbuhler A, Miller RA. A new approach to the implementation of direct care-provider order entry. Proc AMIA Annu Fall Symp. 1996:689–693.
18. Stead WW, Borden R, Bourne J, et al. The Vanderbilt University fast track to IAIMS: transition from planning to implementation. J Am Med Inform Assoc. 1996;3(5):308–17.
19. Clayton PD, Sideli RV, Sengupta S. Open architecture and integrated information at Columbia-Presbyterian Medical Center. MD Comput. 1992;9:297–303.
20. Huff SM, Haug PJ, Stevens LE, Dupont RC, Pryor TA. HELP, the next generation: a new client–server architecture. Proc Annu Symp Comput Appl Med Care. 1994:271–5.
21. Humphreys BL, Lindberg DAB, Schoolman HM, Barnett GO. The Unified Medical Language System: an informatics research collaboration. J Am Med Inform Assoc. 1998;5:1–11.
22. McGrath S, Parseme 1st: SGML for Software Developers.

Englewood Cliffs, NJ: Prentice Hall, 1998.
23. Extensible Markup Language (XML) Web site. Available at: http:llwww.w3.org/XML. Accessed Dec 13, 1999.
24. Geissbuhler A, Miller RA. Distributed knowledge maintenance for clinical decision-support systems: the ''knowledge library'' model. Proc AMIA Annu Symp. 1999;23:770–4.
25. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. J Am Med Inform Assoc. 1994; 1:35–50.
26. Musen MA, Schreiber AT. Architectures for intelligent systems based on reusable components. Artif Intell Med. 1995; 7:189–99.
27. Musen MA. Dimensions of knowledge sharing and reuse. Comput Biomed Res. 1992;25:435–67.
28. Uschold M, Gruninger M. Ontologies: principles, methods, and applications. the knowledge engineering review. 1996; 11:93–136.
29. Musen MA. Domain ontologies in software engineering: use of protégé with the EON architecture. Methods Inf Med. 1998;37:540–50.
30. Winograd T, Flores F. Understanding Computers and Cognition: A New Foundation for Design. Norwood, NJ: Ablex, 1986.
31. Miller RA, McNeill MA, Challinor SM, Masarie FE Jr, Myers JD. The INTERNIST-1/Quick Medical Reference Project: status report. West J Med. 1986;145:816–22.
32. Musen MA, Tu SW, Das AK, Shahar Y. EON: a component-based approach to automation of protocol-directed therapy. J Am Med Inform Assoc. 1996;3:367–88.
33. eCo Framework Working Group (1999). eCo Specification Web site. Sep 30, 1999. Available at: http://216.38.137.215/specs/index.cfm. Accessed Dec 13, 1999.
34. Mitchell T. Machine learning and data mining. Commun Assoc Comput Machinery. 1999;42:30–6.
35. The Dartmouth Atlas of Health Care in the United States. Chicago: AHA Press, 1998.
36. Walker M, Blum R. Towards automated discovery from clinical databases: the RADIX Project. Medinfo. 1986:32–6.
37. Jollis J, Ancukiewicz M, DeLong E, Pryor D, Muhlbaier M, Mark D. Discordance of databases designed for claims payment versus clinical information systems: implications for outcomes research. Ann Intern Med. 1993;119:844–50.
38. Cimino J. Desiderata for controlled medical vocabularies in the 21st century. Methods Inf Med. 1998;37:394–403.
39. Hripcsak G, Friedman C, Anderson P, DuMouchel W, Johnson S, Clayton P. Unlocking clinical data from narrative reports: a study of natural language processing. Ann Intern Med. 1995;122:681–8.
40. Ziman J. Information, communication, knowledge. Nature. 1969;224:318–24.
41. Hersh W, Hickman D. How well do physicians use electronic information retrieval systems? A framework for investigation and review of the literature. J Am Med Inform Assoc. 1998;280:1347–52.
42. Dickersin K, Scherer R, Lefebvrec. Identifying relevant studies for systematic reviews. BMJ. 1994;309:1286–91.
43. Boguski M, Chakravarti A, Gibbs R, Green E, Myers R. The end of the beginning: the race to begin human genome sequencing. Genome Res. 1996;6:771–2.
44. Sweeney L. Guaranteeing anonymity when sharing medical data: the Datafly System. Proc AMIA Annu Fall Symp. 1997: 51–5.
45. Mulrow C. The medical review article: state of the science. Ann Intern Med. 1987;106:485–8.
46. Silberg W, Lundberg G, Musacchio R. Assessing, controlling, and assuring the quality of medical information on the Internet: caveat lector et viewor—let the reader and viewer beware. JAMA. 1997;277–5.
47. Hersh W, Gorman P, Sacherek L. Applicability and quality of information for answering clinical questions on the Web. JAMA. 1998;280:1307–8.
48. Hersh W, Brown K, Donohoe L, Campbell E, Horacek A. CliniWeb: managing clinical information on the World Wide Web. J Am Med Inform Assoc. 1996;3:273–80.
49. Price S, Hersh W. Filtering Web pages for quality indicators: an empirical approach to finding high quality consumer health information on the World Wide Web. Proc AMIA Annu Symp. 1999:911–5.