Technologies for Extracting Full Value from the Electronic Patient Record

HCCIS01.pdf

Walter Panko PhD
School of Biomedical & Health Information Sciences, College of Health & Human Development Sciences, University of Illinois at Chicago, MC 530
Chicago, IL 60612
wpanko@uic.edu

Jonathan Silverstein MD
School of Biomedical & Health Information, Sciences, College of Health & Human Development Sciences and Department of Surgery, College of Medicine, University of Illinois at Chicago
jsilver@uic.edu

Thomas Lincoln MD
School of Biomedical and Health Information Sciences College of Health and Human Development Sciences
University of Illinois at Chicago
lincoln@rand.org

Abstract

Unstructured medical texts, such as the reports and narrative portions of electronic patient record, pose numerous problems in data communication, data base navigation, information retrieval, and records management. The computational complexity introduced by these problems stand in the way of physicians who wish to use these electronic modalities to effectively deliver care to their patients and access the knowledge sources that will keep their professional skills current. Three potentially complementary technologies have begun to demonstrate their promise for dealing with these problems: Markup Languages, Concept Spaces, and JAVA. Each requires investing a certain amount of preprocessing or pre-programming up front. We discuss these three technologies separately, but with the partially tested understanding that each can support and leverage the other.

1. Introduction

Of the problems associated with unstructured narrative in the electronic patient record (EPR), narrative processing remains the most acute limitation toward the development of fully effective clinical systems.[1,2,3] Although a recently deployed new generation of systems based on client-server architectures has begun to capture and manage patient information more effectively, there are other strong incentives for continued turnover and change. The new priorities of managed care and innovative approaches to prospective reimbursement have found present systems wanting, due to their narrow focus on in-patient administration alone; meanwhile the timetable for system replacement has been further accelerated by the "Year 2000 Problem" that afflicts many older hospital information systems, making them cheaper to replace rather than to repair.

There is now general agreement that clinical data itself should belong to the institution that collects and manages it, and not to the vendor who provides the tools to do so. Further, these data should not be altered to fit the technology, but rather the technology should accommodate clinical documentation in full. [4] The new emphasis of today's EPRs on clinical data and record keeping is timely, and has been welcomed by administrators and many physicians. However, broad experience with professionals has also shown that it is not enough to demonstrate that information can be delivered; the process must be rapid and easy and well-focused, or it will not be widely accepted. True clinical usefulness remains severely limited by the fact that much of the information of greatest value lies in the difficult to process narrative portion. Just as in their paper chart equivalents, the information in progress notes and other dictated or written reports remain largely inaccessible to most computer-based retrieval and analysis methods. Moreover, the slow development of natural language processing suggests that if one were to rely on this strategy, such

0-7695-0001-3/99 $10.00 (c) 1999 IEEE

inaccessibility would continue for some time to come. The reasons are several: (1) text material in its native form is too amorphous and ambiguous to be put into subsets and rearranged effectively; (2) once components are identified, grouping and indexing them for retrieval so that the results both encompass the query domain effectively and prioritize with specificity has proved to be an even deeper problem; and (3) in a distributed world, where computing is done on a wide variety of equipment using many different data system conventions, complete sharing cannot be done without processing approaches that are independent of these constraints. Here we provide an overview of three emerging information technologies: *Markup Languages, Concept Spaces, and JAVA*. Each one has the potential to mitigate some of these problems.

The technology closest to the direct management of text is the extension of Markup Languages to new uses. These are best known through the emergent Website use of HTML (Hyper-Text Markup Language), which, in addition to its convenient display properties, has demonstrated a powerful navigation capability for exploring links among heterogeneous data bases to make the Internet and the World Wide Web a truly revolutionary phenomenon.[5] Much more disciplined and powerful is SGML (Standard Generalized Markup Language) and its new and vigorous subset XML (Extended Markup Language). Due to XML's recent widespread acceptance as a data description language in many venues of the computer community, it is now uniquely positioned to become a basis for clinical use, eg, XML is capable of describing the contents of the semi-structured text found in most clinical reports, which must also preserve their unity and integrity as signed documents.[6,7] A second, less well-known emerging technology that addresses the navigation and management of large bodies of texts is *Concept Mapping*. Based on statistical methods developed at the University of Illinois at Urbana-Champaign, a powerful type of meta-data is derived by pre-processing texts to extract their concept related semantics.[8] In our domain, this approach promises to automate a logical cross-indexing of diverse clinical narratives using clustered relationships as well as to automate the associated literature in our enlarging libraries. The third technology is the architecture-independent *JAVA* programming language that, by virtue of its independence, can interactively bring the other two technologies to any computer and to any user interface. *Markup Languages* and *JAVA* are already intimately intertwined to carry out numerous generic tasks on the Web.[9] By extending meta-data beyond coded and key word indexing, and beyond inverted files based on free text search, eg, the search engines on the Internet, *Concept Mapping* addresses a missing piece. Together, *Markup Languages*, *Concept Mapping*, and *JAVA* can be expected to facilitate inter-operability among diverse clinical information sources.

2. Concept Spaces

A major barrier to searching literature or other documents lies in the diversity of the language used in different subspecialties and by different authors. This is true of all disciplines. In medicine, as in other fields, mounting frustrations have driven attempts to standardize key words and coding schemes, and to build associative vocabulary networks. [10] Each attempt has been a partial success, but each has also tested the hypothesis that scientists and professionals in any discipline speak (or should speak) a single, universal language. Peter Galison, in a careful study of the language of physicists, has demonstrated quite well how and why this hypothesis fails.[11] He describes how different subcultures within physics swap information with each other at the border of their subspecialties by developing " pidgeons" and "Creoles" that form forgiving "trading logics."

In many respects, the differences between communication styles are deeper than dialects. Indeed, this struggle with the many facets of person to person communication is by no means new. Among the best of the early formulations is that of Philip Wegener who, in 1885, observed that, "All discourse involves two elements, . . . the context (verbal or practical) and the novelty. The novelty is what the speaker is trying to point out or to express. For this purpose he will use any word that serves him. The word may be apt, or it may be ambiguous, or even new; the context, seen or stated, modifies it and determines just what is meant." [12] This linguistic complexity is both a strength and weakness: It allows language to act as a bridge as well as a specifier so that it may move almost effortlessly with the times, contrasting sharply with fixed vocabularies and indices.

Working from such considerations, and with funding from the Digital Library Initiative of the National Science Foundation, Bruce R. Schatz (Director of the Digital Library Research Program at the University of Illinois at Urbana-Champaign in Illinois) and Hsinchun Chen (Associate Professor in

the Department of Management Information Systems at the University of Arizona) have been performing analyses of the frequency of words in documents and the local co-occurrences of words in "text neighborhoods."[13] Using classical methods for inferring meaning from a statistical analysis of word frequency rather than sentence structure (as in natural language processing), they hypothesize that such local co-occurrences (expressed in matrices) indicate a probable link to the same underlying concepts. When implemented on a large scale, the method results in the generation of a large number of "concept spaces" with numerous associative links resulting in both high fidelity and broad coverage. This method was first developed to link vocabulary terms between different, but related intellectual domains. For example, engineers who design suspension bridges and those who design offshore oil platforms encounter many of the same problems in fluid dynamics but use vastly different words and phrases to describe the same underlying physical constructs. The automated translation of terminology by linked association to concept spaces makes it possible for engineers from different domains to examine each other's literature. This is a particularly sophisticated example of the addition of *meta-data* to a database to facilitate navigation and data extraction. The value and validity of this particular approach has been demonstrated using 5,000,000 abstracts from an engineering bibliographic database (INSPEC). Here the investment in *meta-data* was a major computational task of concept extraction and linking, which required the exclusive use of a Convex Exemplar supercomputer for 10 days. Extensive processing was invested in populating an overall information structure, making any subsequent specific search rapid, convenient, and close to the mark. This method and its results intrigued our group at the University of Illinois at Chicago because quite similar issues occur frequently in health care information. Preliminary work with MEDLINE abstracts has demonstrated that the technique may be useful for linking the varied terms used by different health care specialists and workers to describe the same medical concepts; and specifically for extracting sets of similar concepts from the unstructured narrative of the EPR. Following results in other fields, we believe this should facilitate the discovery and association of similar clinical cases; a task well beyond the capability of highly structured (and often artificial) administrative databases.

Additionally intriguing was the potential ability to link the concepts in the published medical literature to those in a patient's record. Making this link can facilitate a relatively new approach to medical teaching and practice termed *evidence-based care*. Here, using a combination of navigation tools, the data in original articles and clinical information can be brought to bear on the solution of particular clinical problems by taking advantage of comparable meta-data from both sources. Using such literature data rather than summary reviews was considered by Rennels in 1988, but the lack of technological solutions--such as those described in this report--severely limited his efforts to very tightly focused clinical problems and the assignment of articles to well-defined relational data models.[14] On the other hand, it is the outliers that cause clinicians the most trouble. It seems that most patients, perhaps 85% or better, cluster about the mean for a given condition, allowing the use of guidelines and standard procedures. However, the more difficult clinical problems generally have multiple facets due to the number of systems involved. In these circumstances, it becomes almost immediately apparent that the unique reliance on a expected physiological (or pathophysiological) mean, or specific anatomy (or histology), is inadequate for quality care. Individual variation is simply too great, and in many instances, critically important. The anticipation of exceptions--and the skill in managing them when they appear—not only identifies a first rate clinician but also a first rate computer-based processing environment for clinical material. When faced with an exceptional case, a clinician must avoid blindly (and expensively) testing for every possible contingency. An automated, concept-based digital library service could augment a clinician's abilities to rationally relate the findings in the literature to the specifics of an individual patient with attention to the novelty and the surrounding context. Application of these services would allow clinical records to resemble well-crafted case reports and case series that give full attention to individual detail, where required, rather than the spectrum of offhand notes to strict scientific notebooks that they resemble today. Bringing such tools to the bedside could rapidly erode the distinction between the clinical literature (including practice guidelines and published case series) and the well-documented clinical record.

The medico-legal and political fallout from visible dissatisfactions with managed care introduces further incentives to bring charts and the literature closer together. Pushed by an alert media, there is growing pressure both in the federal government and in the

state courts toward using malpractice laws to hold managed care firms responsible for patient outcomes. Any dispassionate examination of modern medical misadventures will not only note the unavoidable vagaries of practice, but also the temptations for profit by restricting care and the deficiencies caused by a lack of available knowledge. In risk management, accurate documentation offers concrete advantages for a proper defense (in the course of a tort challenge) and also provides the necessary data to avoid a repetition of an error (where an error has been made). Such new pressures may well make documenting the use of specific knowledge resources in clinical decisions more than an abstract academic notion. For example, anesthesiology, which is a particularly transactional domain with a high level of operative risk, has benefited from accurate documentation, which is reflected in their malpractice insurance. One way that other clinicians can show that a medical decision was made on a sound clinical basis rather than for economic reasons is to document, within the EPR, the knowledge resources consulted before that decision was made. Moreover, being able to point to the clinical literature most relevant to a particular patient makes both documentation and justification easier. Modern clinical information systems (utilizing the suggested technologies) can easily accommodate this task.

3. Markup Languages

A rule-based approach toward dealing with the narrative within an EPR is to add structure to the text by means of content-related tagging. This can be done in a systematic manner that does not force static categories on the information at-hand (as a specific data model would). SGML and XML both provide just such a flexible meta-structure for the description of content. With origins in word processing and publishing, the markup language of record, SGML, has had an eleven year history as an International Standards Organization standard.[15,16,17] With the purpose of formatting a text for printing, e.g., a book, manual, or document, digitized text streams are marked up using commands in brackets, which is similar to troff, LaTex, and the Rich Text Format (RTF). (RTF allows the translation between Word and WordPerfect in their various versions and will shortly be replaced by XML.) [18,19] Such an approach facilitates in-line text changes and simplifies both the editing process and the creation of multiple versions for different audiences. It is now in widespread use. For example, the *New England Journal of Medicine* [20] is set up for publishing in SGML (as are all Elsevier publications), the *Journal of the American Medical Association* is creating electronic versions of its journals according to this convention, and the *National Library of Medicine* is positioning itself for broad on-line retrieval using SGML.[21] However, the greatest initial impetus was provided by the Department of Defense (DOD),which evaluated SGML for the management of its extensive and complex contract documents. Based on that evaluation, DOD required that all contracts be written using SGML. Being quite generalizable, the process was applied to other kinds of documentation, such as the management of the engineering documents at Boeing, and the computer documentation at SUN Microsystems and IBM. In response to such demands, there arose an industry to design tools and applications for document processing that are *SGML aware*. The strength of this approach lies in the ability to introduce an outline into a document of any size, and to refine the outline to any needed depth, using tags and a directed graph of the tag relationships. In a very real sense, the usefulness of the hierarchical data structure that characterized classical data bases has been indirectly reintroduced as a guiding principle for documents stored as flat files of text, but without the constraints of a data base architecture. The directed graph, termed a Document Type Definition (or DTD) facilitates the automated navigation of the text by following the tag descriptors, which are used almost as street signs. In this way, an application can find, extract, manipulate, and perhaps alter a component of particular interest. This straightforward structure can enrich the introduction of other tagged entities that can define category-dependent attributes and pointers to various distant reference points. However, in a manner characteristic of the DOD, by requiring SGML to meet all contingencies as a (nearly static) International Standards Organization standard, SGML was forced into a bloated state that proved awkward to understand in its entirety and expensive to implement. Nevertheless, from a management perspective, it allowed a systematic integration of otherwise heterogeneous material within a single explicit schema. A DTD could describe relationships among the tags that would permit the design of flexible, self-documenting text packages.

This flexibility exceeded expectations. It turned out that applications aware of the SGML conventions could parse and integrate data from diverse systems without a prior common design, and prepare

document subsets from the data for multiple, even unanticipated purposes. It remained for HTML, a simplified *ad hoc* markup designed to support a graphic presentation interface, and for the World Wide Web to demonstrate the full generic capacity of this approach: displaying Web pages and using hypertext to branch among them. Since 1993, the explosive growth of browser technologies and Websites is the best possible testimony for the broad practicality of markup technologies. This development has been particularly spectacular because in a client-server configuration a browser following one set of formatting conventions can be made to fit nearly all presentation contingencies. The additional leverage provided by the *JAVA* processing language has animated the interaction (which is addressed later in the this text).

The use of a disciplined markup language differs from HTML in four important ways.

1. SGML and XML tags introduce no specific actions (like <blink>) or format descriptions (like <B> for bold);
2. the entire focus is on content description, avoiding format description;
3. the tag names (<example>) and how they are used are locally defined by a DTD and are variable, not fixed, eg, it is possible to introduce a domain specific descriptive tag such as "diagnosis"; and
4. importantly, the DTD and element definitions associated with a document or document class can be tested for logical consistency using a compiler so that data misinterpretations or ambiguities are avoided.

Adoption of a disciplined subset of SGML retaining the characteristics that differ from promotes an open standard and protects specific implementations and associated applications from the kinds of competitive volatility presently typical of HTML "browser wars." XML is such a fully disciplined subset of SGML, specifically designed for the Web. It has the power and modular form of SGML with the outward simplicity of HTML. In less than one year from its initial announcement, this carefully simplified set of conventions has received the enthusiastic approval of all major hardware and software vendors and Internet integrators. As a consequence, XML is moving rapidly out of its infancy to become a major player in communication, transaction processing, and data archiving. For example, Microsoft, in their forthcoming generation of office application products, has made XML the hidden data description language, retaining HTML for the display. In this new manner, their applications will become entirely Web-compliant.

What can a markup language do within an EPR? On paper, the natural unit of a clinical chart is a document, generated by someone directly involved in the clinical activities thus recorded who must attest to the document's accuracy by signing it. Such documents take on many forms and cover many different issues. A major problem associated with an EPR based on data elements rather than documents is the loss of both context and integrity when such elements are extracted and isolated from the original report. The "processable" extracted material, best expressed in phrases, is force fit into more limited fields. If large chunks are left as free text instead, they remain difficult to navigate and parse. Moreover, a hard copy must be retained as the legal record; otherwise the signature over the entire content is lost. Tagging can overcome this by offering a set of conventions that can break up the specifics according to a common outline, while retaining the whole. Indeed, in health care reports, it is the outlines that offer the greatest stability, with sub-outlines nested inside the general outline to handle increased specialization; for example, the history and physical will have different content in the hands of a generalist and a specialist. Indeed, the more specialized the practice, the greater the depth and diversity can be found in the context and caveats these professionals consider important, and thus in the tagging required for their characterization. This hierarchy of content is an appropriate markup structure from which object oriented encapsulations of greater focus on particular forms of processing can be extracted.

In July 1997, a working group consisting of the HL-7 Special Interest Group (a special interest group attached to the HL-7 standards committee), individuals from the SGML community, and members of the document processing industry met at the Kona Hotel, in New Hampshire, Rhode Island. The group created an initial framework as a proof of concept for content markup in health care, which became the *Kona Proposal*.[7] The group's intent was to ascertain if *sample* health care documents *heavily marked up* could be *usefully* parsed by the numerous SGML application packages available on the market. This *Kona Proposal* is, in fact, a proposal only, and continues to evolve. It applies a structure that lays out the content of health care reports following the familiar pattern of increasingly explicit detail. The

most general (and forgiving) level merely labels the document as a whole with an identifying header (the *Kona Header*) according to type, eg, admission note, history and physical; progress note, OP-note, discharge note, etc. The second level (provisionally termed the SOAP level) uses a general formal outline beginning with subjective input, objective results, assessment of the data overall, and a plan with particular steps. This follows a convention widely applied in medicine, often in abbreviated form. It can also be used to describe the contents of reports of any type of consultation. (Drawn up by nonsurgical medical specialists, the description of surgical actions was not covered in this initial draft. The surgeon's plan is to act, which requires its own description.) A third level is designed include both the designated specialized novelty together with certain additional modifying contexts. A fourth level was initially proposed to accommodate local concerns such as an audit, fine-structure that identifies times, individuals, hand-off, etc., associated with particular actions (usually only of internal importance in a particular site, but capable of documenting chain of custody, and the like).

When documents were drawn up for test purposes according to the first two levels of this formulation, it was found that these could be almost effortlessly manipulated by the products and applications of more than 20 vendors at a conference titled "the XML Mixer" held in La Jolla, California, later in July 1997. For data manipulation and display, only the most minor adjustments were necessary in tagging the commonest everyday reports, thus presenting not only a proof of concept, but also a strong demonstration of potential practical usefulness.

In establishing an overall structure, one does not escape the need for a meta-description of each general document type. However, the potential for absorbing the inevitable variations of individual documents under a common architecture is very simplifying. Furthermore, new words to label content are not sought for each kind of document. Instead, available vocabularies and categories from the various coding methodologies and other meta-document descriptors are expected to populate the tagging schemas as content descriptor tags, or as modifying attributes and indices. Moreover, in using this convention, different coding schemes do not compete with each other for a single code slot, but can be introduced simultaneously through separate tags in series, so long as room is made for them in the meta-schema: the DTD. This permits interpretation by a wide range of vocabularies based on the retained, original text.

Following roughly on the *Kona Proposal* philosophy of nested detail, we have found that prolonged narratives like an operative report can be effectively marked up, beginning with a raw operative report. Prior work by Galen Cook in the late 1960s outlined how this might be done without the tools to do it. [22] The question still remains: How much detail should be identified for future use? Statistical studies to extract conceptual constructs based on concept mapping are also in order to investigate the descriptive requirements of a particular document.

Documents loosely (or semi-) structured by markup are clearly more flexible than records divided into fully specified fields.[23] In communication between various health care sites, data streams between databases presuppose certain conceptual models of clinical care. In the classical HL-7 exchange, a complete agreement with respect to the positioning of each field is required. These conventions have served well within a given organization; however, these data models become less useful as the scope of the exchange is enlarged. Different operational perspectives among services and institutions lead to conceptual models that become more pleomorphic and volatile as the scope enlarges. The same data elements could be extracted from both the present HL7 data stream and from templates, but these latter semi-structured units promise to be more useful with respect to post-processing.[24]

Health care data processing has always been conservative in adopting new directions. Is there any reason to believe that the use of markup languages will become either relevant or widespread in the near term? The best reason is the wide acceptance of the new XML standard by the networking and network server community at large. Following Willy Sutton's dictum, "go where the money is," their initial applications are directed toward business: electronic commerce and financial *push technologies* that can bring together requested data from different heterogeneous sources for decision making using the Web. This general computer leadership certainly accounts for the recent flurry of activity between the HL-7 standards and the SGML community with XML taking the lead as the most likely candidate to deal easily and adequately with the complexity and volatility of clinical information. Barriers remain, not the least of which is the momentum of prior design activities involving many players and vendors. It is hard to give up a long-held development direction.

However, Bill Gates was able to turn on a dime and embrace browsers and the Web once he saw the power of the technology. Health care, in one manner or another, will do the same. Indeed, members of the International health care community have already moved aggressively to introduce XML as pilot projects within ongoing production work in Sweden, Switzerland, Germany, and the United Kingdom, with additional studies in Japan, Italy, and France.[25]

4. JAVA

JAVA is a portable object-oriented programming language that is executed in a virtual machine that can run on top of nearly any operating system or hardware. If conceptual maps and descriptive markup are passive, *JAVA* is active and is a processing language of choice for XML. Released by Sun Microsystems, Inc., in 1995, it arose from a project of James Gosling and associates. Shortly thereafter Netscape Corporation licensed *JAVA*. Its impact on the Internet and on computing vendors has been profound. Its promise of "write once, run anywhere" has been a holy grail for developers for some time. From the beginning, the *JAVA* team designed it so that applications written in *JAVA* will run on computers from mainframes to the computer chips embedded in appliances, such as a set-top box for cable service or a palmtop computer. While *JAVA* can be used to write applications that run on a computer, it is also adaptable to a form (called applets) and as *JAVA* Beans, which can be downloaded and executed by WWW browsers.

We bring *JAVA* into this discussion because such a language is key to the effective use of the technologies described in the previous two sections.[6] Applications must act on marked up data or conceptual maps in order to have any output of interest. *JAVA* applets permit the development of inexpensive *thin clients* that are suitable for ubiquitous use in health care. *Thin client* is the term given to computer platforms (hardware and software) that hold a minimal set of software and data locally, but also get their software and data from network servers as needed. (Once again the world is profoundly changed.) In addition to reducing the support costs for distributed computing by centralizing the availability of the latest software version, *thin clients* also allow one or more users to work with multiple devices at many locations with one set of data and applications. *Thin clients* may be network computers or even dedicated *JAVA* computers. The most common *thin client* architecture today is one that uses simple and inexpensive distributed computers to support WWW browsers with *JAVA* applets. These same browsers are the programs that can use and interpret XML.

Why are *thin clients* and portable *JAVA* applets so important to managing information in EPRs? The answer is *scaling*. *Scaling* is a term used for computer-based systems that function on a scale different from the development or current environment. Until recently, and still predominantly today, *scaling* was a concept that took the number of users from a limited number to a very large number, ie, *scaling up*. However, in health care informatics today, there is a growing concern about *scaling down*, ie, moving EPR access and use from a resource-rich and sophisticated academic medical center onto systems that can be used and maintained in a one- or two-physician practice.

It is clear that today's emerging modern and sophisticated clinical information systems are not ideal for *scaling down*. Their use would impose unacceptably large support and financial burdens. In addition, the level of service and reliability would also likely suffer. *Thin clients*, implemented on simple and inexpensive computers, offer the potential for efficient and cost-effective access to EPRs. Through the use of WWW browsers, the *thin clients* can use XML and the results of the concept matching to bring effective EPR access to a small practice or a mobile health care professional. Thus, *JAVA*'s role in tapping the potential of *Concept Spaces* and *Markup Languages* is likely to be crucial. If it isn't *JAVA*, then it will be some other similar technology.

5. Discussion

Our analysis reveals that many of the problems associated with unstructured text in the clinical narrative can be addressed with pre-processing, open architectures, flexible data exchange standards, multiple views, and data-models that retain context.

Just as in the preparation for concept searching, tagging requires an investment. This can now be minimized by using SGML- and XML-aware (Markup Languages) word processors that prompt data entry into a flexible outline. Alternately, interactive encounter forms, screens, or check-off lists that generate tags in the course of creating the documents (e.g., pen-activated check-off screens) can interactively suggest common phrases as actual content responses and--properly programmed--can

minimize effort while leaving the user free to introduce their own text.

Current clinical information systems will not be of much help, even if they adopt the concept mapping and tagging technologies. The complexity and size of these systems makes their ubiquitous deployment prohibitively expensive. By enabling the use of simplified *thin clients*, most of the capabilities of clinical information systems can be *scaled down* to accommodate small, technology-sparse environments. The ubiquitous use of these technologies will certainly enhance quality of care as well as further justify the investment in the pre-processing.

Most importantly, *Markup Languages* and *Concept Spaces*--properly designed using open architectures and tagging schemes--are stakeholder-neutral so that they may be used by different systems (eg, those that are physician, nursing, administratively oriented) in multiple and often unanticipated ways. Even where a need for context is minimal, such semi-structured documents offer what may be the optimal way of passing information between heterogeneous information systems in a truly technology- and vendor-independent manner.

Used for communication, markup languages are demonstrably more flexible than the current data exchange standard and are explicitly designed through the introduction of an indirect "meta-description" to anticipate and deal with future changes in data structure and significance. Similarly, the generation of concept spaces (co-occurrence maps) of a set of documents uses only the documents themselves as input data. This automatically ensures that the concept spaces slowly shift with the times, thereby adjusting the data-model as understanding and significance shift.

Because markup-structured documents and *Concept Spaces* can incorporate multiple vocabularies simultaneously, these techniques can be used universally for all health care documents (including the literature) in an integration toward *evidence-based care* where the harvesting and cross correlation of data, adjusted for individual differences, is critical to the success. Additional vocabulary references are conceived as overlays, retaining the original version of the document in the data, while not interfering with any type of security or version identification required. One could describe these as a new form of "data-model," which would exist to add context, not to structure the content *per se*.

In the matter of coded indexing, any number of separate codes can be introduced as attribute tags to a given textual component such as a problem, diagnosis, or procedure. Fixed vocabularies can be introduced as indices in the same manner, allowing both individual variation in description and disciplined wording for navigation. A chosen tag set is an example of one fixed vocabulary.

Ultimately, technologies which permit the processing of data context as well as the immediate content will permit the wide range of human expression to be viewed, when needed, and specific statistical data to be extracted appropriately. By further developing these technologies with platform independence, virtually unlimited possibilities emerge for communication within the EPR.

6. Conclusions

The true potential of the EPR will not be realized until the limitations of unstructured text in the clinical narrative are appreciated and overcome. We believe the three technologies discussed here: *Markup Languages*, *Concept Spaces*, and *Java* offer significant potential for progress in this area. The use of *Concept Spaces* could greatly facilitate the association of the highly specific notations in patients' records with other records and with the clinical literature. By interactively suggesting the most productive search terms, *Concept Spaces* can potentially reduce the overwhelming task of applying patient care databases and the medical literature to a particular problem. Similarly, *Markup Languages* coupled with architecture-independent programming languages can generate multiple-use documents and data models out of a single source. There is a strong potential for synergy between these three methods. Our group will continue to work with all of them to learn more about the nature of, and solutions to, the unstructured text problem. Since the Internet and intranets are rapidly becoming the basic framework of clinical computing, these three Web-based technologies can be expected to have an important impact beyond the boundaries of any single institution, reaching into the domains of integrated health care delivery networks and community health information networks.

References

1. Lincoln TL, Essin DJ, Ware WH. The electronic medical record: A challenge for computer science to develop clinically and socially relevant computer systems to

coordinate information for patient care and analysis. *The Information Society*. 1993;9:157-188.
2. Hara S. Construction of the infrastructure for health expert system. In: Ifeachor EC, Rosen KG, eds. *Proceedings of the International Conference on Neural Networks and Expert Systems in Medicine and Healthcare*; 1994 August 2323-26; Plymouth, United Kingdom.
3. Lincoln TL, Essin DJ, Anderson RD, Willis H. The introduction of a new document processing paradigm into health care computing-A CAIT [white paper]. Available at: www.mcis.duke.edu/standards/SGML/ proposals/CAIT-white-paper.txt. Accessed May 29, 1998.
4. Alschuler L, Dolin R, Spinosa J. In: The next decade--pushing the envelope. Proceedings of *the SGML 1997 Conference*; 1997 May 11-15; Barcelona, Spain. p. 1561-1565.
5. Lindberg D, Humphreys AB, Betsy L. Medical informatics. *JAMA*. 1996;23:275.
6. Lincoln TL. Codifying medical records in XML: Philosophy and engineering. In: Khare R, Connolly D, eds. *XML: Principles, Tools, and Techniques* [serial online]. Fall 1997;2(4):149-152. Available from: OReilly & Associates, Sebastopol, California. Accessed Fall 1997.
7. Radosevich, L. Health care uses XML for records [Kona Proposal]. August 25, 1997. Available at: www.infoworld.com/archives/html/97-i01-34.51.htm. Accessed May 29, 1998.
8. Schatz BR. Information retrieval in digital libraries: Bringing search to the net. *SCIENCE*. 1997;275:327-334.
9. Bosak J. XML, Java, and the future of the Web. In: Khare R, Connolly D, eds. *XML: Principles, Tools, and Techniques* [serial online]. Fall 1997;2(4):219-227. Available from: OReilly & Associates, Sebastopol, California. Accessed Fall 1997.
10. Cimino JJ. Auditing the Unified Medical Language System with semantic methods. *J Am Med Inform Assoc*. January 1998:5(1): 41-51.
11. Galison, P. *Image and Logic*. Chicago, Illinois: Chicago University Press; 1997.
12. Wegener, P. Untersuchungen uber die Grundfragen des Sprachlebens [reissued by Konrad Koerner]. Philadelphia, Pennsylvania: Benjamin Publishing Co.; 1991.
13. Schatz B, Chen H. [University of Illinois at Urbana-Champaign *DeLIver* Website.] Available at: dli.grainger.uiuc.edu/. Accessed May 29, 1998.
14. Rennels GD. *A Computational Model of Reasoning from the Clinical Literature*. Ann Arbor, Michigan: University Microfilms International; 1986.
15. Alschuler L. *ABCD... SGML: A Users Guide to Structured Information*. London, England/Boston, Massachusetts: International Thomson Computer Press (ITCP);1995.
16. Bray T, Paoli J, Sperberg-McQueen CM. Extensible markup language (XML). In: Khare R, Connolly D, eds. *XML: Principles, Tools, and Techniques* [serial online]. Fall 1997;2(4):67-82. Available from: OReilly & Associates, Sebastopol, California *and* www.sil.org/sgml/sgml.html. Accessed Fall 1997.
17. Goldfarb CF. *The SGML Handbook*. Oxford, England: Oxford University Press; 1990.
18. Edwards M. XML: Data the way you want it [Microsoft Website]. October 31, 1997. Available at: www.microsoft.com/sitebuilder/workshop/author/xml/xmldata-f.htm. Accessed May 29, 1998. (Microsoft policy statement with respect to open architecture, followed by a demonstration of the same at SGML/XML Europe 1998 conference. May 24, 1998. Available at: www.sil.org/sgml/edwards971104.html. Accessed May 29, 1998.
19. Paoli J, Schach D, Lovett C, Layman A, Cseri I. Building XML Parsers for Microsofts IE4. In: Khare R, Connolly D, eds. *XML: Principles, Tools, and Techniques* [serial online]. Fall 1997;2(4):187-195. Available from: OReilly & Associates, Sebastopol, California *and* www.sil.org/sgml/sgml.html. Accessed Fall 1997.
20. Sabasteanski A. Use of the electronic manuscript standard at the *New England Journal of Medicine*. EPSIG News. March 1989;2(1):1,2.
21. Data Conversion Laboratory. [National Library of Medicine Website.] Making medical information available on-line. Data Conversion Laboratory (DCL) research report. September 1996. Available at: www.dclab.com/nlm.htm. Accessed May 29, 1989.
22. Cook GB, Watson FR. Linear graphing of surgical decisions and activities. *J Surg Res*. June 1969;9(6):361-367.
23. Williams JP. Healthcare informatics standards: An electronic health record developers perspective [working paper]. Available at: www.hytime.org/ihc97/papers/williams.html. Accessed May 29, 1998.
24. Dolin RH, Alschuler L, Bray T, Mattison JE. SGML as a message interchange format in healthcare. Presentation at *The 1997 American Medical Informatics Association (AMIA) Annual Fall Symposium*; 1997 October 25-29; Nashville, Tennessee.
25. Dudeck J. HL-7 user group in Germany. Presentation in conjunction with *the SGML/XML 1998 Europe conference*; 1998 May 19-22; Giessen, Germany.