Revolutionary impact of XML on biomedical information interoperability

A. Shabo
S. Rabinovici-Cohen
P. Vortman

The use of Extensible Markup Language (XML) to implement data sharing and semantic interoperability in healthcare and life sciences has become ubiquitous in recent years. Because in many areas there was no preexisting data format, XML has been readily embraced and is having a great impact. Biomedical data is very heterogeneous, varying from administrative information to clinical data, and recently to genomic data, making information exchange a great challenge. In particular, it is hard to achieve semantic interoperability among disparate and dispersed systems—a common constellation in the fragmented world of healthcare. Moreover, the emerging patient-centric and information-based medicine approach is posing another challenge—the development and use of an integrated health record for each patient. This means that diverse data from many systems has to be generated, integrated, and become available at the point of care. This paper presents the case that XML is becoming the integration "glue" for biomedical information interoperability, which can lead to improvements in pharmaceuticals, genomic-based clinical research, and personalized medicine, which, for the first time, can be fine-tuned to serve individuals through their longitudinal electronic health records.

INTRODUCTION

The use of Extensible Markup Language (XML) in Healthcare and Life Sciences (HCLS) is spreading rapidly with the effort to integrate biomedical information and develop semantic interoperability among the numerous and heterogeneous systems currently in use.[1–3] These disparate but interrelated systems range from those in biological and clinical research laboratories to those at the point of patient care. XML is a natural choice, as it facilitates the creation of self-describing, platform- or application-independent text, and thus provides the crucial infrastructure for information technology applications targeted at semantic interoperability. Because XML tags are based on natural language, biomedical data expressed in XML can be both read by humans and processed by machines. Although it is true that end users are not likely to read raw XML files

without dedicated viewers, it is still quite common for people such as implementors, testers, and knowledge-acquisition and management experts to create or scrutinize XML documents.

Proprietary data formats are a fundamental barrier to semantic interoperability. A decade ago we could still see the preponderance of data represented in proprietary formats at both the technical and semantic levels. At the technical level of physical formats, we could see all variants of formats, from databases to text files to binary files to application-specific files. At the semantic level, there were a few attempts to standardize codes used to represent basic phenomena. For example, in healthcare, we could see the standardization of codes for laboratory results,[4] diseases,[5] medications,[6] and procedures.[7] Similarly, in genomics there have been attempts to create standard identifiers for genes, mutations, and so forth.[8]

Nevertheless, the semantic interoperability challenge is much more complex than just those basic codes. It is about the ability of a system to operate on data from other systems as easily as it operates on its own data. This involves both static and dynamic data representations. Static representations deal with data models of such constructs as clinical documents and genetic testing results. Dynamic representations deal with interactions among systems (carrying static data as message payloads) and the workflows that must be fulfilled.[9] In the past, workflows were processed mainly within the enterprise and primarily served enterprise requirements. Recently, we see a growing demand for cross-enterprise capabilities focused on consumer needs.[10] In HCLS, health consumers are typically patients facing certain medical conditions that disturb their normal life. A patient-centric approach in a world of growing mobility requires a greater degree of semantic interoperability among the various enterprises that provide patient care, whether it is done directly in healthcare or indirectly by the life science enterprises whose efforts lead to new drugs, treatments, and therapeutic devices.

XML technology is also playing a crucial role in the realization of the emerging concept of the electronic health record (EHR). While healthcare providers hold the medical and patient records they create and maintain, the EHR is intended to be a longitudinal, cross-institutional, individual-centric information entity, possibly spanning an individual's lifetime. In addition to medical data, it could also comprise personal information, such as lifestyle, self-documentation, and work environment. In such a long-term effort, there is a need to integrate into one framework data from a variety of sources and based on different models, terminologies, languages, and practices (e.g., clinical guidelines). The objective would be to maintain a coherent EHR that is always available where and when a patient requests care—at any caregiving facility around the globe. This is a huge semantic challenge, whose resolution could be assisted by an XML infrastructure. At the moment, there is no agreed-upon EHR information standard, but efforts are being made to standardize the definitions of functions fulfilled by EHR systems, i.e., systems that manage health records.

A decade ago, the XML format was well-situated to serve the complex needs of HCLS; it had emerged as a means to integrate elements from the various points along the HCLS continuum. However, this *glue* (which is how XML is commonly described because of its usefulness in integration) is only at the technical level, and, verbose as they might be, XML tags cannot provide semantic interoperability without standards for the meaning of tags and their relationships. Nonetheless, XML still represents a crucial step toward semantic interoperability because it provides the common infrastructure on which semantics can be easily standardized and conveyed in real implementations.

In the next three sections, we describe how XML is used in clinical data, clinical-trial data, and genomic data. These three types of data cover the main three domains of the HCLS realm. We emphasize the key standards used to drive semantic interoperability and how XML is used to represent the data in each one of these domains. We then describe the efforts to integrate data across the three domains and describe how such integration is enabled by XML. Finally, we describe the XML technology challenges posed by HCLS that are driving the requirements for an efficient and powerful storage and query mechanism.

XML IN CLINICAL DATA

XML could be better used in healthcare if specific XML structures were standardized as canonical formats to send such data as laboratory results and prescriptions. Health Level Seven (HL7)[11] is the

major standards-developing organization in healthcare, and its Version 2 (V2) specifications have been ubiquitous in many hospitals in the United States and Europe for more than a decade. Nevertheless, the V2 common representation is an ASCII (American Standard Code for Information Interchange) format in which data values are separated by delimiters (known as the *bar format*). ***Figure 1*** shows a typical example of an HL7 message carrying laboratory results. The example clearly shows the inflexibility and unreadability of this format; for example, if a value is missing, it is still required to insert the delimiters, and the meaning of existing values is not always clear.

```
MSH|^~\&|||||19941122100053||ORU^M01|
EVN|M01|199411181141|
PID|||661041||GARDNER^REED^M|
PV1||I|E7^703^^LDS|
OBR||^A000520|LYTES^Serum Electrolytes|
OBX|1|NM|NAS^Serum Sodium|1|138|mmol/L|
OBX|2|NM|K^Serum Potassium|1|3.2|mmol/L|
OBX|3|NM|CL^Serum Chloride|1|114|mmol/L|
OBX|4|NM|CO2^Serum CO2|1|24|mmol/L|
```

Figure 1
A typical HL7 laboratory results message

For the past few years, HL7 has been developing V3, which is being built around a central Reference Information Model (RIM). It has four core classes: Entity, Role, Participation, and Act, which are interrelated based on a Unified Service Action Model.[12] The V3 RIM is the main differentiator between V3 and V2. It is meant to assure consistency throughout various specifications developed by the HL7 technical committees and thus support semantic interoperability in healthcare. The RIM is built on dedicated vocabularies and data types. All V3 specifications must be derived from the RIM. In this way, V3 achieves a higher level of semantic consistency throughout the various specifications. The V3 specifications are balloted (American National Standards Institute [ANSI]-mandated procedures to assure consensus-based development of standards) in their model formats and only then translated to XML by what is called in HL7 *Implementation Technology Specification* (ITS). The latter represents the objective of keeping the balloted specifications technology-neutral, and allowing different ITSes to translate the models to different implementation technologies. At the moment, only the XML-ITS exists for V3 as a normative specification.[13]

The Structured Documents Technical Committee of HL7 has been a strong advocate of moving to XML representation.[14] Their Clinical Document Architecture (CDA) standard was the first XML-based standard approved by ANSI in 2000. The second release of the CDA standard[15] has been recently approved, and ***Figure 2*** shows a portion of a sample instance supplemented with the standard. The XML shows the use of constructs like `observation`, with its nesting tags, such as `code` and `value`, to represent an observation. These tags represent major observation class attributes: `id` holds a unique identifier of this observation, the code holds the type of observation, and the value holds the observation itself. The complete CDA XML sample can be found within the standard specification package.[11]

Other standards organizations, such as the ASTM[16] and IHE,[17] develop standards that are partially based on XML representations. For example, ASTM develops a continuity-of-care-record (CCR) format,[18] which is a summative referral letter used to refer a patient from one healthcare provider to another, thus assuring continuity of care. It includes a summary of the patient's health status (e.g., problems, medications, and allergies) and basic information about insurance, advance directives, care documentation, and care plan recommendations. An attempt is made to represent the CCR format as a set of constraints over the generic HL7 CDA standard by using XML-related languages, like XPath.

XML IN CLINICAL TRIALS

In the pharmaceutical industry, the clinical research cycle is both long and expensive; to develop a single drug can take up to several years, and costs have soared to around $800 million per drug. The purpose of the clinical research cycle is to test how well new medical treatments or other interventions work in people. It includes numerous areas of activity, such as protocol development, enrollment of subjects, collection and processing of data, clinical trial management, data analysis, and submission to a regulatory agency. These activities are performed by diverse systems that are required to share data extensively. Use of XML for the interchange of clinical-trial data could eliminate manual transcription of data from one system to another.

```
<observation classCode="OBS" moodCode="EVN">
  <id root="10.23.4573.15879"/>
  <code code="313193002" codeSystem="2.16.840.1.113883.6.96"
    codeSystemName="SNOMED CT" displayName="Peak flow"/>
  <effectiveTime value="20000407"/>
  <value xsi:type="RTO_PQ_PQ">
    <numerator value="260" unit="l"/>
    <denominator value="1" unit="min"/>
  </value>
</observation>
```

Figure 2
A portion of a CDA sample showing an observation of the peak flow of a patient's lungs

Moreover, because XML includes both data and metadata in a universal format, namely text, the data can be correctly represented at its destination point, regardless of the platform or application used to create it and no matter how technologies may evolve in the future.

The Clinical Data Interchange Standards Consortium (CDISC) is leading the development of standards to improve data quality and accelerate product development in the pharmaceutical industry.[19] The CDISC model focuses on the use of metadata, and the approach is to combine XML representation with the tabular presentation traditionally used for clinical-trial data. One of the CDISC standards is the Operational Data Model (ODM)—an XML-based representation to solve the problem of moving data from any collection system to the central database of the sponsor of the clinical trial. ODM includes clinical study metadata (e.g., item definitions and protocols), clinical study administrative data (e.g., users and access privileges), and clinical study data (e.g., complete records of patient data and audit trail). It does not specify specific items for clinical trials, but rather furnishes a container that includes a definition of grouped items along with the actual data that complies with that definition.

Another important CDISC standard is the Study Data Tabulation Model (SDTM)—a tabular format used to submit data to regulatory authorities, such as the United States Food and Drug Administration (FDA). The model can be deduced from the ODM data or from other transport file formats, such as an .xpt SAS** transport file. Along with SDTM, the CDISC specifies `Define.xml`—an extension of ODM that includes metadata on the domains and variables used in SDTM. ***Figure 3*** is an excerpt from `Define.xml` that includes the metadata of 'te.xpt', where the latter is represented in the traditional tabular format used for FDA submissions. The metadata is presented in elements such as `def:Label`, which classifies the contents of the tabular file, and the element `def:DomainKeys`, which describes the key attributes of that file.

One additional important element is `xlink:href`, which includes the link to the referenced tabular file. CDISC is also working with HL7 to harmonize its standards and extend the clinical trial standards from the HL7 V3 RIM.

XML IN GENOMIC DATA

Breakthroughs in genetic sciences hold promises for new personalized medications and treatments, improved clinical care, and refined preventive medicine. Exploiting these promises to their full potential depends on integrating and correlating the vast amount of dispersed clinical data, clinical-trial data, and genomic data—a process in which XML is playing a key role. Previously, the bioinformatics arena did not have established non-XML data formats, such as those that existed for healthcare and clinical trials; therefore, the adoption of XML to create shared representations of genomic data was more readily accepted. A vast amount of genomic data has been evolving around the Human Genome Project, in which all human chromosomes were sequenced to reveal the genetic code and how it is divided into genes.

The Bioinformatic Sequence Markup Language (BSML) was the first XML application in the life sciences to represent sequences data.[20] This language was developed by LabBook Software under a grant from the National Human Genome Research

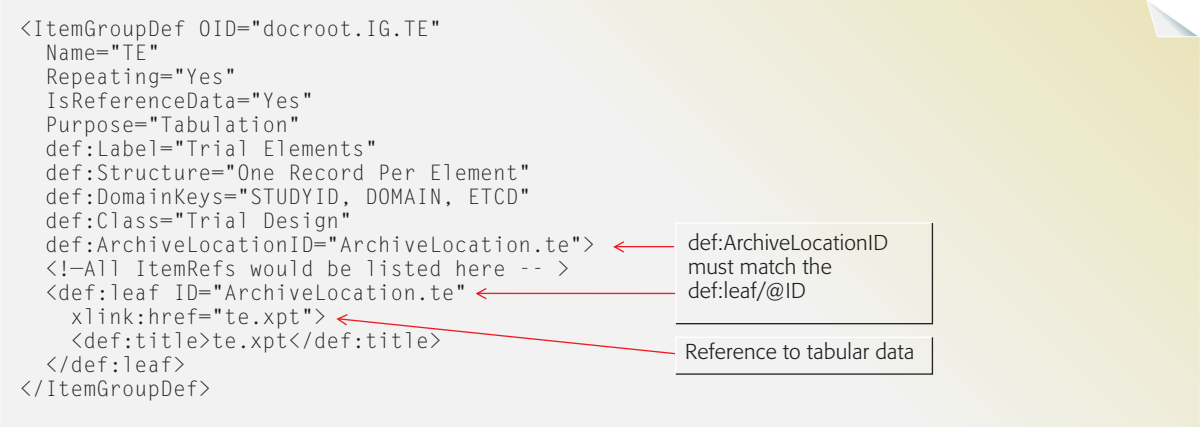

Figure 3
Portion of an ODM Define.xml sample showing reference to traditional tabular format

Institute to provide standard methods for communicating genomic research information. BSML mixes graphical and display representation with actual data, which overloads the XML files and is sometimes inconvenient.

Some of the genomic information resides in public repositories for data storage and mining, and most of those resources can export their data in XML. One of the largest public databases of this sort is the National Center for Biotechnology Information (NCBI), which disseminates biomedical information such as gene sequences, single nucleotide polymorphisms (SNPs), protein sequences, and medical publications. NCBI can export its data in XML. Similarly, the European Bioinformatic Institute (EBI) offers a service to display data from the European Molecular Biology Laboratory nucleotide sequence databank in XML format. However, the XML formats of those two organizations do not comply with the same schema; therefore, they facilitate interoperability only partially. To have true semantic interoperability, the various XML schemas must be harmonized in a single model and standardized.

Standardization has proved successful when it comes to representing gene expression data. The Microarray Gene Expression (MAGE) Group and the Object Management Group (OMG) standard organizations are developing the MAGE standard for exchanging microarray data generated by functional genomics and proteomics experiments.[21] MAGE describes microarray designs, microarray manufacturing information, microarray experiment setup and execution information, gene expression data, and data analysis results. The EBI ArrayExpress[22] database is a public resource of gene expression data represented in MAGE.

Genomic data representation is especially challenging because the high-throughput genomic tests generally produce a large amount of data that needs to be represented as one unit. Although XML is not economic in the size it uses for data representation, it was still chosen by various organizations as the representation format. This is mainly due to the interoperability that XML facilitates, its semistructured nature, its human readability, and its processability by applications. Indeed, MAGE files may be very large—some of them can reach up to several gigabytes in size. These files include several hundred samples of various specimens, including thousands of genes that were tested under numerous conditions. Additionally, MAGE files sometimes include links to external files, which may in turn be in either XML or tabular format, making the data units even larger. Thus, the storing, processing, and querying of MAGE files requires special consideration.

Another important role of XML—one that started with genomic data but is now also used in other domains—is providing the glue infrastructure for Model Driven Architecture** (MDA**). MDA assumes that domain knowledge is expressed in models, so that tools and services semiautomatically

translate those models to technology-specific components. Because HCLS is an emerging field with rapidly evolving requirements, there is a need for the semiautomatic generation of flexible technology components that can meet these requirements. Because MDA separates domain knowledge from the underlying technology specifics, each technology component can evolve at its own pace. The current MDA practice relies on XML to transfer the domain knowledge expressed in models to the technology-specific components, and thus is an important enabler of MDA. The National Cancer Institute has been particularly active in promoting MDA with genomic data through its caBIG** project[23] and related cancer research initiatives.

XML IN CLINICAL GENOMICS

In the previous sections we presented the use of XML in clinical and clinical-trial data representations and in bioinformatics. As genomic knowledge is being accumulated rapidly and reliable links to clinical practice are established, it is important to cope with the challenge of integrating genomic, clinical, and clinical-trial data related to the same person with a coherent representation. Several HL7 organizational members (e.g., IBM, Mayo Clinic, Cap Gemini, Ernst & Young) have initiated the Clinical Genomics Special Interest Group (SIG), whose scope is the actual use of genomic data in healthcare practice.

The main specification developed by the HL7 Clinical Genomics SIG is the Genotype model. The Genotype model is intended to be used as a shared component in any HL7 specification that conveys genomic data. It embeds various types of genomic data relating to a chromosomal locus, including sequence variations, gene expression, and proteomics. The Genotype model utilizes the existing bioinformatics markups commonly used by the genomic community (e.g., MAGE Markup Language for gene expression data or BSML for DNA sequences). The bioinformatics markups represent the raw genomic data and are encapsulated in HL7 objects. On the other hand, only portions of the raw and mass genomic data are relevant to clinical practice. To that end, the full-blown bioinformatics markup schemas have been constrained, and areas describing pure research data were excluded. More important, the Genotype model also includes specialized HL7 objects (e.g., `SequenceVariation` of SNP type) that hold those portions of the raw genomic data that are significant to clinical practice. These specialized objects have attributes that represent the essential genomic data along with optional annotation. They are populated through a bubbling-up process carried out by dedicated applications (*bubbling up* means computing processes that analyze massive raw data and bring to the surface specific data items relevant to the use case at stake).

■ Because XML tags are based on natural language, biomedical data expressed in XML can be both read by humans and processsed by machines ■

This process should take into account the goals of clinical care, the patient-specific history, and the most current knowledge about relevant clinical-genomic correlations. Once populated, these specialized objects can be associated with clinical phenotypes, represented either internally within the Genotype model or elsewhere (for example, as diagnoses, allergies, and adverse drug events residing in patient medical records). Savel et al.[24] used the Genotype model to represent C-reactive protein, pentraxin-related data in the National Health and Nutrition Examination Survey project of the United States Centers for Disease Control and Prevention.[25] A few XML exemplary instances were generated to demonstrate the usability of the Genotype model.

Figure 4 is an overview of the Genotype model, highlighting the key classes in the model. Note the entry point at the top (any genetic locus) and the clinical phenotypes at the bottom right, to which all genomic classes are connected.

The main paradigm underlying the design of this standard can be described as *encapsulate and bubble up*, and it aims at addressing the coexistence of clinical data formats with markups for bioinformatics. The encapsulation phase is a static process in which certain encapsulating objects in the Genotype model are populated with portions of raw genomic data based on predefined constrained bioinformatics markups. The constraining process is part of the standardization effort and is designed to leave out portions that seem irrelevant to the clinical practice; for example, the display elements in the BSML markup. The constraining process also makes

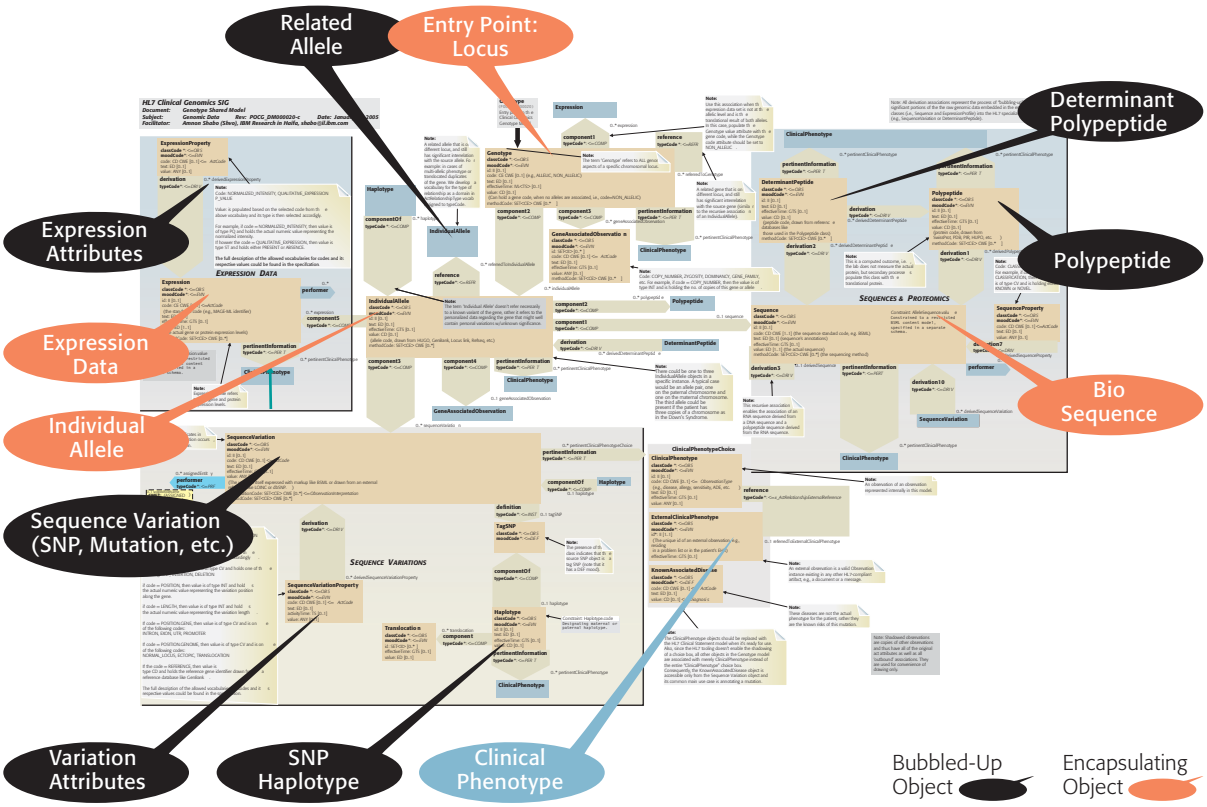

Figure 4
Overview of the Genotype model; callouts point to key classes in the model, which are associated with each other using ActRelationship association classes

sure that the data refers to one patient and one gene only to fit the Genotype scope. To represent multiple gene data, allelic (alternative form) data, and locus data, higher-level models using the Genotype model are available (e.g., Genetic Profile, Tissue Typing, and Family History models).

The bubbling-up phase is a dynamic process in which genomic-oriented decision-support applications parse the raw genomic data encapsulated in the HL7 instance and identify and select those portions that seem to be clinically significant to the patient's clinical history and treatment goals, based on the most updated scientific knowledge. The results of this bubbling-up process are held in other HL7 objects in the Genotype model. These objects can also be associated with clinical data in the patient's medical records (represented in the Genotype model as the Clinical Phenotype classes).

These static and dynamic phases lead to a gradual distillation of the raw genomic data in the context of diagnosis and treatment provided to a specific patient, while holding the parts of the raw data within the HL7 objects so that they can be parsed again when new knowledge becomes available. The complete raw genomic data will be accessible only by reference, possibly by using the Life Sciences Identifier, a new OMG specification.[26]

Clinical genomics data integration requires complex workflows, such as the encapsulate and bubble up paradigm explored by the HL7 Clinical Genomics SIG. ***Figure 5*** shows a workflow in which the above coexistence of biological and medical data takes place and is executed stepwise. First, the static phase takes place, and HL7 messages are sent to an EHR system with encapsulating objects carrying raw genomic data. In the second phase, these messages are enriched with bubbled-up objects that are required by the end-user application at the point of care. The figure shows an example taken from the sequencing type of data: the most clinically significant SNPs are bubbled up from the raw sequencing data and associated with clinical phenotypes.

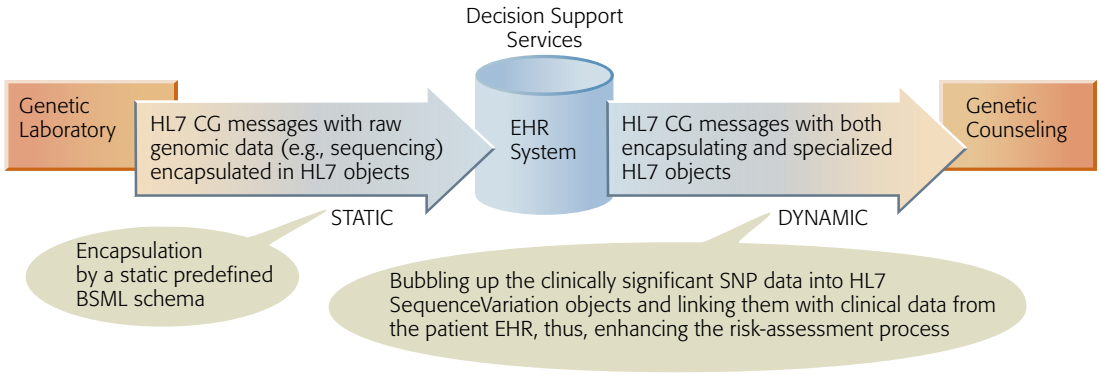

Figure 5
Clinical genomics (CG) workflow: encapsulate and bubble-up (exemplified with sequencing data)

Note that several bubbling-up processes could be performed at the same time (e.g., different algorithms or ontologies) and at different times (e.g., when new discoveries become available and the same raw data should be reinterpreted). Therefore, it is important not to lose the raw genomic data of a specific patient to abstraction; rather, encapsulate it and make it available to any processes that attempt to associate it with clinical data to facilitate clinical decisions at the point of care.

Complex workflows and computations are needed to implement an approach such as encapsulate and bubble up. Content models from totally different worlds must be subtly integrated. Because XML is so effective as the glue among these worlds, it represents a revolutionary solution to this challenge. A concrete example is presented in ***Figure 6***: the need to exchange family history data when risk assessment is required for cancer patients.

Figure 6 is a concrete example of how clinical and genomic data are brought together when family history data must be exchanged for risk assessment of cancer patients, as demonstrated through an implementation[27] of an HL7 family history model where the Genotype model is utilized. An XML schema was automatically generated from the model by using the HL7 XML-ITS (see the section "XML in clinical data" for more details). This made it possible to create XML samples with an actual patient's family history data which validate against that schema.

The fragments shown in Figure 6 are taken from a family history sample composed in XML that represents the clinical and genomic data of a patient who has a mother and a father (each of whom has two parents), two sisters, a husband, and a daughter. (The full sample is available from the HL7 site.[11]) The XML instance starts with the patient as the root element (Figure 6A). The fragment in Figure 6B describes the patient's daughter, who died of breast cancer. The genomic data appears first, identifying a specific allele of the BRCA2 gene.

The fragment in Figure 6C shows an elaboration of the BRCA2 allele by encapsulating sequences from that allele that might consist of personal SNPs beyond those variations that typically identify this allele. (Note that the DNA sequences are presented for illustration purposes only and are not necessarily authentic.)

XML TECHNICAL ISSUES

Although XML technologies positively impact the HCLS domain, the latter poses additional technical challenges for XML technology. XML schemas (and the XML Schema Definition language) allow services and clients running on diverse platforms to interoperate over a common type set and are critical for the successful use of XML. The schemas used in HCLS are quite complex, with a high level of nesting, recursion, compound data types, and dynamic typing (where the type of some XML data fields is known only at runtime). Moreover, the XML instances may be very large and may include structured as well as unstructured data. A discharge summary, for example, includes patient name, gender, age, and so forth, which is structured data, but it also contains free text description of the hospitalization course, which is unstructured data. The CDA standard[15] supports this combination by having specific elements for structured data and a

A

```
<Patient xmlns="urn:hl7-org:v3" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="urn:hl7-org:v3POCG_MT004008.xsd">
   <id extension="555.001-SUBJ"/>
   <id extension="555.002-NMTH"/>
   <id extension="555.003-NFTH"/>
   <!-- PATIENT-->
   <patientPerson>
     <administrativeGenderCode code="F"/>
     <birthTime value="1957"/>
     ...
```

B

```
   <!-- DAUGHTER-->
  <relationshipHolder>
     <id extension="555.011-SUBJ"/>
     <id extension="555.001-NMTH"/>
     <id extension="555.01-NFTH"/>
     <code code="DAU"/>
     <relationshipHolder>
       <deceasedInd value="true"/>
     </relationshipHolder>
       <!-- GENOMIC DATA-->
     <subjectOf>
       <Genotype>
         <component2>
           <individualAllele>
             <text>breast cancer 2, early onset</text>
             <value code="U43746" displayName="BRCA2" codeSystemName
               ="HUGO"/>
     ...
```

C

```
<sequence>
  <code code="BSMLcon3"/>
    <value>
      <Definitions>
        <Sequences>
          <Sequence id="seq1" molecule="dna"
            ic-acckey="U14680 REGION: 101..199"
            db-source="GenBank" title="BRCA1, exon 2" representation="raw"
            local-acckey="this could be used by the genetic lab">
            <Seq-data>
              GCTCCCA CTCCATGAGG TATTTCTTCA
              CATCCGTGTC CCGGCCCGGC CGCGGGGAGC CCCGCTTCAT
              CGCCGTGGGC TACGTGGACG ACACGCAGTT CGTGCGGTTC
              GACAGCGACG CCGCGAGCCA GAGGATGGAG CCGCGGGCGC
              CGTGGATAGA GCAGGAGGGG CCGGAGTATT GGGACCAGGA
              GACACGGAAT GTGAAGGCCC AGTCACAGAC TGACCGAGTG
              GACCTGGGGA CCCTGCGCGG CTACTACAAC CAGAGCGAGG
              CCG
            </Seq-data>
          </Sequence>
          ...
```

Figure 6
Family history XML sample: (A) root element fragment, (B) daughter fragment, and (C) sequencing fragment

text element for narrative data. The structured and unstructured parts may or may not describe the same piece of information, and the structured part may have anchors to the narrative part and vice versa.

HCLS is driving the requirements for efficient and powerful XML storage and query mechanisms. Structured data has a definite mathematical model (relational calculus) and a powerful query language (Structured Query Language) whose implementations have been optimized over many years. Similarly, unstructured data has a sound information retrieval model and a fast flexible search mechanism coupled with ranking algorithms to retrieve data according to its relevance. However,

the semistructured data typically found in XML files has neither a defined storage model nor a preferred query or search language. There is no specialized storage model for XML to handle collection, relationship, metadata, updates, binary data, and

■ Use of XML for the interchange of clinical-trial data could eliminate manual transcription of data from one system to another ■

XML data that references structured data. Given the lack of a mature and powerful storage model specialized for XML, XML files are either treated as structured data and stored in a relational database or treated as unstructured data and stored in a content management system.

In the first case, the XML files are shredded (decomposed) into a relational database, mainly when data mining and business intelligence need to be obtained from that data. Although most database products support shredding of XML files, they have usually failed to carry it out for the HCLS standards because of the complexity of those schemas (e.g., HL7 CDA). Today, however, the shredding tools are advancing to meet HCLS requirements.

In the case in which the XML files are treated as mostly unstructured data containers, the XML files are stored in a content management system, but the queries for these documents are generally not powerful enough for HCLS requirements. For example, it is difficult to find patients who have two diseases, where each disease appears in a different document (known as a *join operation* in a relational database).

We are now witnessing the early emergence of native XML databases, which are storage models specialized for XML. Native XML databases can generally store document collections of various schemas and are thus more flexible than the relational model. But, these efforts are still experimental. Performance needs to be improved and query and search capabilities need to be enhanced. One such experiment is XML File System (XMLFS), which provides a file system interface to navigate XML repositories.[28] It uses a specialized information retrieval index to manipulate data values in the context of their elements and attributes. On another front, the new Java** Content Repository specification also promises to bridge the gap between the structured and unstructured worlds and to provide a good storage model for XML data.[29] These emerging technologies still need to reach a higher level of maturity before they can be used in the HCLS domain.

Querying a collection of documents that include structured and unstructured data is challenging. XPath is sometimes the query language of choice, but it lacks some important capabilities, such as cross-document joins. XQuery is more comprehensive and robust when it comes to parametric search and navigation within XML documents, but it lacks text search and navigation among documents and from collection to collection. The XQuery Full-Text language is expected to provide the additional free text search required.[30]

Another area that needs active research to improve XML adoption is the handling of large XML instances, such as the gene expression MAGE files. Not only are the units of genomic data very large, but XML is verbose and consumes a large amount of space for the amount of actual data being sent. Performance is a concern when handling XML data. To address this problem, the World Wide Web Consortium is investigating binary serialization formats of XML instead of the standard textual representation.[31] Binary XML may reduce network bandwidth when XML streams are transferred, and it can reduce the time and space required for parsing and processing XML files. There are various methods for creating binary XML files, and they are currently being evaluated (for an example, see Reference 32). However, it is still not clear whether only one binary encoding standard can satisfy the needs of all applications or whether binary encoding will result in incompatible versions of XML.

CONCLUSION

The ultimate goal in HCLS is to base healthcare practice on clinical research and clinical trials in a way that patient treatment is improved, patient safety is enhanced, and clinical research is facilitated in an ongoing iterative cycle of development. XML is supporting these processes by being the common data representation infrastructure that

facilitates biomedical information integration and interoperability within and across the fragmented world of healthcare, along with the pharmaceutical and bioinformatics industries.

Because XML lacks semantics at its core, semantics is provided by standards. In this paper, we have reviewed several major emerging XML-based standards used with clinical data, clinical-trial data, genomic data, and combined clinical-genomic data. We have also discussed how the standards are represented by XML and which features they exploit. It seems that the complexity of the HCLS domain drives the requirements provided to the XML technology community to new levels.

Semantic interoperability in HCLS, technically enabled by XML, is still in its early stages. To achieve that goal, the XML-based standardization efforts must continue intensively and include the use of ontologies, low-level data representation schemas, and workflows. To increase the adoption of those standards, XML technology must improve its performance, handle large files and complex schemas, enable treatment of structured and unstructured data as if it were of the same kind, and provide efficient and powerful storage, processing, and query and search mechanisms that can be generated using the MDA approach. IBM Clinical Genomics Solutions—stemmed from the Secure Health and Medical Access Network (SHAMAN) research project—makes use of many of the XML-based standards mentioned in this paper, for example, in the iCAPTURE[33] and the Ste-Justine Hospital[34] implementations.

CITED REFERENCES

1. J. J. Berman and K. Bhatia, "Biomedical Data Integration: Using XML to Link Clinical and Research Data Sets," *Expert Review of Molecular Diagnostics* **5**, No. 3, 329–336 (2005).
2. P. A. Covitz, "To Infinity, and Beyond: Uniting the Galaxy of Biological Data," *Omics—A Journal of Integrative Biology* **7**, No. 1, 21–22 (2003).
3. R. Schweiger, S. Hoelzer, U. Altmann, J. Rieger, and J. Dudeck, "Plug-and-Play XML—A Health Care Perspective," *Journal of the American Medical Informatics Association* **9**, No. 1, 37–48 (2002).
4. Logical Observation Identifiers Names and Codes (LOINC), Regenstrief Institute, Inc., http://www.regenstrief.org/loinc/.
5. International Classification of Diseases (ICD), World Health Organization, http://www.who.int/classifications/icd/en/.
6. RxNorm, Standard Names for Clinical Drugs, U.S. National Library of Medicine, http://www.nlm.nih.gov/research/umls/rxnorm/.
7. Current Procedural Terminology (CPT) of the American Medical Association, http://www.ama-assn.org/ama/pub/category/3113.html.
8. NCBI at a Glance, National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/About/glance/ourmission.html.
9. K. Anyanwu, A. Sheth, J. Cardoso, J. Miller, and K. Kochut, "Healthcare Enterprise Process Development and Integration," *Journal of Research and Practice in Information Technology* **35**, No. 2, 83–98 (May 2003).
10. W. A. Yasnoff, B. L. Humphreys, J. M. Overhage, D. E. Detmer, P. F. Brennan, R. W. Morris, B. Middleton, D. W. Bates, and J. P. Fanning, "A Consensus Action Agenda for Achieving the National Health Information Infrastructure," *Journal of the American Medical Informatics Association* **11**, No. 4, 332–338 (2004).
11. Health Level Seven (HL7), an American National Standards Institute (ANSI)-accredited Standards Developing Organization operating in the health-care arena, http://www.hl7.org/.
12. G. Schadow, D. Russler, C. Mead, J. Case, and C. McDonald, "USAM—Unified Service Action Model—Documentation for the Clinical Area of the HL7 Reference Information Model," Regenstrief Institute for Health Care Publication, *Regenstrief Institute for Health Care*, Indiana University School of Medicine, Indianapolis, IN 46202 (2000).
13. E. W. Huang, D. W. Wang, and D. M. Liou, "Development of a Deterministic XML Schema by Resolving Structure Ambiguity of HL7 Messages," *Computer Methods and Programs in Biomedicine* **80**, No. 1, 1–15 (2005).
14. L. Alschuler, SGML Initiative in Health Care (HL7 Health Level-7 and SGML/XML), http://xml.coverpages.org/gen-apps.html#HL7-SGML.
15. R. H. Dolin, L. Alschuler, S. Boyer, C. Beebe, F. M. Behlen, P. V. Biron, and A. Shabo, "HL7 Clinical Document Architecture, Release 2," *Journal of the American Medical Informatics Association* **13**, No. 7, 30–39 (2006).
16. American Society for Testing and Materials (ASTM) International Standards Worldwide, http://www.astm.org.
17. Integrating the Healthcare Enterprise (IHE), Healthcare Information and Management Systems Society, Chicago, IL 60611, http://www.himss.org/ASP/topics_ihe.asp.
18. *Continuity of Care Record (CCR), The Concept Paper of the CCR, Version 2.1b*, American Society for Testing and Materials (ASTM) International Standards Worldwide, http://www.astm.org/COMMIT/E31_ConceptPaper.doc.
19. Clinical Data Interchange Standards Consortium (CDISC), http://www.cdisc.org.
20. Bioinformatic Sequence Markup Language (BSML), http://www.bsml.org.
21. MicroArray and Gene Expression MAGE-ML, http://www.mged.org/Workgroups/MAGE/mage.html.
22. ArrayExpress database, European Molecular Biology Laboratory (EMBL) European Bioinformatics Institute (EBI), http://www.ebi.ac.uk/arrayexpress/.

23. cancer Biomedical Informatics Grid, (caBIG), National Cancer Institute Center for Bioinformatics, https://cabig.nci.nih.gov/overview.
24. T. Savel, B. Lin, A. Shabo, and G. M. McQuillan, "The Future of Transmitting Human Genomic Data in the Public Health Information Network (PHIN): Using a Prototype Health Level 7 Shared Genotype Refined Message Information Model (HL7 R-MIM) and the Extensible Markup Language (XML)," *Proceedings of the 3rd Annual Public Health Information Network (PHIN) Conference*, Atlanta, GA (2005), http://www.cdc.gov/phin/05conference/05-11-05/4G_Savel.pdf.
25. National Health and Nutrition Examination Survey (NHANES), Centers for Disease Control and Prevention (CDC), http://www.cdc.gov/nchs/nhanes.htm.
26. *Life Sciences Identifiers RFP Response*, The Object Management Group, http://www.omg.org/docs/lifesci/03-12-02.doc.
27. A. Shabo and K. S. Hughes, "Family History Information Exchange Services Using HL7 Clinical Genomics Standard Specifications," *International Journal on Semantic Web & Information Systems* **1**, No. 4, 44–67 (2005).
28. A. Azagury, M. E. Factor, Y. S. Maarek, and B. Mandler, "A Novel Navigation Paradigm for XML Repositories," *Journal of the American Society for Information Science and Technology (JASIS)* **53**, No. 6, 515–525 (2002).
29. JSR-000170 Content Repository for Java Technology API (Final Release), Java Community Process Program, Sun Microsystems, http://www.jcp.org/aboutJava/communityprocess/final/jsr170/.
30. S. Amer-Yahia, C. Botev, J. Dörre, and J. Shanmugasundaram, "XQuery Full-Text Extensions Explained," *IBM Systems Journal* **45**, No. 2, 335–352 (2006, this issue).
31. *Report From the W3C Workshop on Binary Interchange of XML Information Item Sets*, World Wide Web Consortium (W3C), http://www.w3.org/2003/08/binary-interchange-workshop/Report.html.
32. R. J. Bayardo, D. Gruhl, V. Josifovski, and J. Myllymaki, "An Evaluation of Binary Encoding Optimizations for Fast Stream Based XML Processing," *Proceedings of the 13th International World Wide Web Conference,* New York, NY (2004), pp. 345–354.
33. *IBM Life Sciences Clinical Genomics Solution Helps iCAPTUR4E Centre Understand Genetic Causes of Disease*, IBM Systems, http://www-306.ibm.com/software/success/cssdb.nsf/CS/RAHE-6E23F3?OpenDocument&Site=eservermain.
34. "A first in Canada: The Ste-Justine Hospital Pediatric Research Centre and IBM Canada join forces to advance research into childhood leukemia—Implementation of an informatics infrastructure designed to help researchers speed treatment and improve patient outcomes," IBM Systems, http://www-03.ibm.com/industries/healthcare/doc/content/news/pressrelease/1340566105.html.

Accepted for publication October 26, 2005.
Published online May 16, 2006.

Amnon Shabo (Shvo)
IBM Haifa Research Laboratory, Haifa University, Mount Carmel, Haifa 31905, Israel (shabo@il.ibm.com). Dr. Shabo is a research staff member and is currently involved in various IBM Healthcare and Life Sciences projects. He is the facilitator of the HL7 Clinical Genomics SIG and its main contributor. In addition, he is a coeditor of CDA Release 2 and a primary editor of its Implementation Guide. He specializes in longitudinal and cross-institutional electronic health records. Dr. Shabo was the visionary and a coauthor of the mEHR (Electronic Health Records for Mobile Citizens) proposal made by a consortium of 19 partners to the European Commission's Sixth Framework Program, based on his vision of independent health record banks.

Simona Rabinovici-Cohen
IBM Haifa Research Laboratory, Haifa University, Mount Carmel, Haifa 31905, Israel (simona@il.ibm.com). Mrs. Cohen is a research staff member. She holds B.S. and M.S. degrees in computer science from the Technion, Israel Institute of Technology. Her areas of interest include information integration and knowledge management systems, especially in the biomedical domain. Mrs. Rabinovici-Cohen is the Haifa project leader of the IBM Clinical Genomics solution.

Pnina Vortman
IBM Haifa Research Laboratory, Haifa University, Mount Carmel, Haifa 31905, Israel (vortman@il.ibm.com). Mrs. Vortman holds an M.S. degree in mathematics and physics from the Hebrew University, Jerusalem. She has been involved in many research projects in a variety of areas and was active in initiating research areas in information-based medicine and integrated health records. She is a main contributor to more than 24 patents and has published over 10 papers. Mrs. Vortman has been a member of the IBM Academy since 1997. ■