Context-Mediated Semantic Alignment of Heterogeneous Electronic Medical Record Systems For Integrated Healthcare Delivery

Guoray Cai

College of Information Sciences and Technology
Penn State University, University Park, PA 16802
Email: cai@ist.psu.edu

Abstract—Healthcare organizations are facing the challenge of integrating their electronic medical records (EMR) for the purpose of unified healthcare service delivery. Through an analysis of the nature of medical data and the use of medical records, we argue that the most significant barrier for service integration is the semantic heterogeneity across fragmented and disintegrated healthcare systems and health records. Based on our understanding of the requirements for EMR semantic integration and the limitations of the existing solutions, we present an approach for modeling the context-dependent nature of medical data semantics. The approach is formalized into a framework (Context-Mediated Semantics Interoperability), which serves as the basis for computational support to medical service integration. Our initial contextualization effort focused on capturing the heterogeneity across different types of specialties, as well as the semantic differentiation for individuals, organizations, communities, and social use.

Keywords-service integration; healthcare delivery; electronic medical records; contextualization; ontology; semantic web;

I. INTRODUCTION

Modern medical service enterprises must deal with integration of information nuggets and silos that exist in different units of health services and medical operations in order to enable meaningful sharing, reuse, and collaboration [3-7]. A huge challenge towards that goal is the semantic barrier of communication and interoperability for diverse use and reuse of heterogeneous medical data [1, 2, 8, 9]. While differences in language, data schemas, data models, and communication protocols can be reconciled through existing solutions (such as standardization, model integration, ontology integration) [10-12], these methods do not work well when semantic heterogeneity presents.

This research seeks deeper understanding of the nature of semantic heterogeneity in electronic medical data and service integration, and proposes a conceptual framework towards meaningful solutions targeting healthcare domain. The US Congress made it to law (HITECH - Health Information Technology for Economic and Clinical Health) Act to provide financial incentives to physicians and other healthcare stakeholders to adopt Electronic Health Records. A core requirement of the Act is to demonstrate “Meaningful Use”[13-16]. One of the foremost barriers is the fragmented and disintegrated healthcare systems and health records. With as many as 20 healthcare providers per patient, a patient’s health record may include a diverse set of views and interpretations of the patient’s health conditions [17]. They often present inconsistent and even conflicting records of diagnosis, treatment plans, and instructions. Effective communication and collaboration among providers is critical to achieve quality and safety outcomes [17, 18]. A recent national analysis of the value of interoperability suggested that fully interoperable healthcare systems could save the nation $77.8 billion annually[19]. Thus, it is a high priority to implement interoperable EMRs across healthcare organizations. Pure ontology-based approaches for interoperability (such as OpenEHR project [5]) have made limited success. This is not surprising, based on the lessons learned from other domains.

As the first step towards a solution, this paper presents a critical review of the literature to establish a conceptual framework for EHR semantic integration. The review is driven by the following questions: (1) what is the nature of ‘medical work’ and what are the roles of the (electronic) medical record? (2) what are the sources and causes of semantic heterogeneity? (3) what are the requirements for HER semantic integration? (4) what are the solutions available and how well they work? We seek answers from a diverse literature in the fields of health informatics, healthcare delivery, information science, and semantic web.

The second part of this paper proposes a framework for addressing the semantic heterogeneity problem in the integration of electronic medical records. The core idea is the extension of existing ontology-based semantics data models to include the contexts of use as a new layer of semantics in the modeling stack. We will describe the framework for semantic interoperability, which is called Context-Mediated Semantics Interoperability (CMSI framework). It specifies a semantic model in terms of three related components: activity-centric context representation [20, 21], contextualized ontology space [2], and context mediated semantic exchange [22]. Refinement of the framework as well as the computational methods will be done through a set of use scenarios taken from two types: heterogeneity across different types of specialties, (2) semantic differentiation for individuals, organizations, communities, and social use.
II. ENTERPRISE INTEGRATION IN HEALTHCARE ORGANIZATIONS

American healthcare system is facing unprecedented challenges and opportunities, and is due for major re-engineering of the whole system. On one hand, health care is substantially underperforming on most dimensions: effectiveness, appropriateness, safety, cost, efficiency, and value. According to recent reports [17, 23],

- U.S. healthcare system is the most expensive healthcare system in the world. Yet it is among the least effective. US healthcare spending was about $7,439 per person and accounted for 16.3 percent of the nation’s gross domestic product (GDP) in 2007 and will trend upward reaching 19.5 percent of GDP in 2017 [23].
- It contributes to only 5 to 10 percent of total health—an estimated 3.5 to 7 years of lifespan—derives from the health care delivery system
- Between 44,000 and 98,000 people die each year from preventable injuries sustained as part of care delivery in U.S. hospitals.
- Waste: According to the National Academies, between 30 and 40 percent of healthcare costs—more than half a trillion dollars per year—is spent on “overuse, underuse, misuse, duplication, system failures, and unnecessary repetition, poor communication, and inefficiency.”

A fundamental cause of these problems is the fragmented nature of the healthcare delivery system. Fragmentation exists in all levels of healthcare system, from organizational, incentive structures, service provisions, and information technology infrastructure [24]. In particular, there is a high degree of fragmentation in treatment and care. A patient in the US is served by as many as 20 healthcare providers in lifetime. A patient’s health record may include a diverse set of views and interpretations of the patient’s health conditions[17]. They often present inconsistent and even conflicting records of diagnosis, treatment plans, and instructions. Fragmentation of treatment and care is the result of increasing specialization in health care that has lead to increasing regimens by specialists. Unfortunately, the holistic focus and interdependence of treatment nuggets tends to be lost.

In a fragmented healthcare delivery system, there is ever stronger the need of integration and synthesis. High-quality care requires the smooth flow of information across diverse providers working within various organizations in both inpatient and outpatient settings. A first step in fixing the system would be to build the interoperable electronic medical records that should improve coordination among providers and reduce gaps in care. A fully realized interoperable healthcare IT system could reduce errors, improve communication, help eliminate redundancy, and provide numerous other benefits that would protect patients and save up to tens of billions of dollars per year.

The central challenge to achieving such a system is interoperability—the ability of data systems, medical devices and software from different vendors based on a diverse array of platforms to share patient EHRs, electronic physician orders for lab tests and drug prescriptions, electronic referrals to specialists, electronic access to information about current recommended treatments and research findings, and other information. The vision, as stated by the “healthcare information enterprise integration initiative” [4] is to transform the healthcare system from isolated treatment episodes towards one that is patient-centric, coordinated, and continuous treatment process involving multiple healthcare professionals and various institutions. A recent national analysis of the value of interoperability suggested that fully interoperable healthcare systems could save the nation $77.8 billion annually [19].

Thus, it is a high priority to implement interoperable EHRs across healthcare organizations. President Obama initiated a massive national HIT program and allocated $20 billion in funds for Health Information Technology for Economic and Clinical Health or HITECH, etc.

III. THE NATURE OF 'MEDICAL WORK' AND THE ROLES OF ELECTRONIC HEALTH RECORDS

Medical work is the combination of diagnosis, treatment, and prevention of disease, illness, injury, and other physical and mental impairments in humans. It involves caregivers who practice medicine, chiropractic, dentistry, nursing, pharmacy, allied health, and other services. For optimal patient care, the various provider organizations and health professionals have to cooperate closely during patient care, often called shared care or integrated care. Shared care is defined as the continuous patient-oriented cooperation of hospitals, general practitioners (GPs), specialists and other health care professionals during patient care [5]. Shared care imposes great challenges on the availability and processing of information including trusting shared information, the correct and clinically safe interpretation of the information.

Electronic Health Records (EHRs) mainly involve clinical patient data including the personal and family history, the clinical state, the dispensed therapies, and other relevant information about the reached diagnostics and outcomes [25]. An EHR is used primarily for purposes of setting objectives and planning patient care, documenting the delivery of care and assessing the outcomes of care. It also includes information regarding patient needs during episodes of care provided by different health care professionals. This includes items like handwritten, typed, or electronic clinical notes; notes recorded from telephone conversations; all correspondence including letters to and from other health care professionals, insurers, patients, family, and others; laboratory reports; radiographs and other imaging records; electrocardiograms and printouts from monitoring equipment; audiovisuals; and other computerized/electronic records, including e-mail messages.

Since the first conception of electronic health records (HER) in the 1990s, the content, structure, and technology of EHR have been evolved, driven by the basic idea of supporting and enhancing health care, and improving service delivery and its quality [26]. In 2009, the American Recovery and Reinvestment Act mandated that hospitals in the USA move to electronic medical records (EMR) systems by 2014. It prescribes the implemented EMR to have the following characteristics: (1) patient-centered; (2) longitudinal—it is a long-term record of care, (3) comprehensive—it includes a record of care events from all types of care givers, providers and institutions tending to a
patient, not just one specialty, and (4) prospective - not only are previous events recorded, but also instructions and prospective information such as plans, goals, orders and evaluations.

An interoperable EMR with the above characteristics will serve as the foundation for translating the vision of evidence-based practice [27] into reality. An integrated EMR system serve as a record of the longitudinal health history of each patient is required to improve quality of care.

IV. THE SEMANTIC HETEROGENEITY OF MEDICAL RECORDS

Building interoperable electronic medical records must start with addressing data heterogeneity among systems and data collections. The proliferation of data heterogeneity among autonomous systems creates islands or silos of medical records, and presents major barriers towards full realization of the financial, clinical, and efficiency benefits conferred by HIT [28].

Crucial dimensions of data heterogeneity are syntactic, structural, and semantic heterogeneities (see Figure 1).

1. **Syntactic heterogeneity**: each system may represent medical data and knowledge using different encoding and communication format semantic interoperability paradigms, such as relational or object oriented models. Currently, there is no single universally accepted clinical data model that will be adhered to by all [29].

Syntactic interoperability ensures that clinicians can always send information to another provider and receive information which they can read.

2. **Schematic heterogeneity**: this refers to lack of interoperability due to the differences in the way each system structure objects and their relationships in medical contents. For example, objects in one system are considered as properties in another, or object classes can have different aggregation or generalization hierarchies, although they might describe the same Real World facts.

3. **Semantic heterogeneity**: Semantics is defined as the meanings of terms and expressions. Hence, semantic interoperability is the ability of information systems to exchange information on the basis of shared, pre-established and negotiated meanings of terms and expressions [5].

In semantically interoperable EMR, different information systems used by the various health care providers of shared care can understand the context and meaning of information provided by other systems. However, there are great variations across information systems in the descriptions of diseases, causes, and treatments, due to the highly specialized and dynamic service landscape.

Today powerful integration tools (e.g. application servers, object brokers, different kinds of message-oriented middleware, schema mapping, and workflow management systems) are available to overcome syntactical and schematic heterogeneity of medical record systems. Yet, semantic heterogeneity remains as a major barrier to seamless integration of patients’ medical records. We will focus on semantic heterogeneity in the rest of this report.

A major contributor to semantic heterogeneity is the high degree of specialized service providers that apply different subset of medical knowledge to patients. Efforts for resolving semantic heterogeneity have taken multiple approaches. The most dominant approach is to establish standards for establishing meanings of medical terminologies and expressions. The development of vocabulary standards to support use of EHRs has been quite successful in the last two decades, mounted by the United States National Library of Medicine in its UMLS project, the UK National Health Service and its Centre for Coding and Classification, by SNOMED-CT International, and by the GALEN program of the European Community. Despite many years of concentrated and coordinated effort to build comprehensive medical terminology standards, a single agreed-upon system of global electronic medical reference terminologies and ontologies does not yet exist by today and there has been doubts if such a goal is possible [2, 30].

While a single global medical ontology does not exist, there are many vocabulary systems that are in use including anatomy concepts, ICD9, ICD-O-3, INDEX VIRUM, LOINC, MEDCIN, Med- DRA, SNOMED and the clinical drug codes being developed by the Veterans’ Administration and the Food and Drug Administration. In addition, the National Library of Medicine defined global electronic medical ontologies [26, 31-33].

Existing terminology systems that are developed in academic research projects are fundamentally flawed from the point of view of practical use in scalable systems, and this explains why commercial vendors of HIT systems rarely choose them. Figure 2 shows the gap that current standardization efforts have not addressed.

V. SEMANTIC HETEROGENEITY OF CLINICAL TERMINOLOGIES

Clinical terminology concerns the meaning, expression, and use of concepts in statements in the medical record or other clinical information system. Healthcare professionals manage constantly with minor differences of meaning and
even misunderstandings of information written by their colleagues or received from other institutions. The ideal solution would be to reach a full consensus on a terminology standard that supports both the practical use by clinicians and computer systems for automated processing and analysis. This seemingly straightforward problem (from semantic web perspective) has been partially proven to be extremely hard in medical information, for many good reasons [32, 34]. This is one of those application domains where “special purpose solutions to small scale problems are easy, while fully scalable general purpose solution is extremely difficult.” Another domain of similar nature is geographic information science which has been articulated by Cai [35]. Figure 3 shows the types of vocabularies that have been developed so far.

Reflecting on the slow progress in developing a comprehensive re-usable terminology for patient-centered systems, Rector [30] put forward ten reasons why clinical terminology for coordinated EMR is hard. I summarize a few of important ones below, since they are fundamental to the arguments I make in this paper.

[Reason 1] It is difficult to scale up terminology solutions to the vast and the multiplicity of activities, tasks and users to be served by EMR. The scope of medical knowledge across all specialties, for detailed clinical care, in particular, is orders of magnitude larger than the terminology needed to report simple diagnostic registers in a single specialty or even general practice. Scaling up by an order of magnitude or more is notoriously difficult, because the complexity of digital systems tends to increase exponentially with the scale of the vocabulary.

[Reason 2] It is difficult to come up with terminology solutions that meet the requirements of usability by both human and machine processing. Humans and machines process information very differently. Humans are compliant, flexible, tolerant, while machines require us to be rigid, fixed, and intolerant. They present fundamentally conflicting requirements that we have no good solutions so far [36].

[Reason 3] It is difficult to come up with terminology solutions when the complexity of clinical pragmatics is introduced. Clinical pragmatics includes three aspects: clinical conventions, clinical expectation, and operational meaning. These pragmatic complexities often defy any attempts to capture them fully in formal representations. Phrases do not literally mean what they say. Same phrases can be interpreted in multiple ways, with some of them more usual than others (for instance ‘Heart Valve’ would seem to mean “valve in the heart”, but physicians may understand it in four different ways, so additional information on common usage beyond either linguistics or the formal concept representation is required). The meaning of a term that is intended to cover may vary in scope with the situation within which an operational record was created, and is often undetermined outside that operational context. On the other hand, it is almost a necessity to embrace clinical pragmatics. A terminology that has no concern and proven relevance clinical pragmatics is useless. Unfortunately, developers of medical terminology have largely ignored the need to explicitly capture the pragmatic knowledge together with the terminology.

Given the above understanding of the requirements for medical terminology, the current solutions are seriously flawed! According to Rector [30] and Cimino [37], academic research has assumed that we can clearly separate linguistic knowledge, the medical concept systems, and pragmatic knowledge associated with clinical terminology.

[Hypothesis of Separability]: For a clinical terminology, the representation of concepts and the relations between them can and should be separated from the linguistic knowledge about how these concepts are expressed in language and the pragmatic knowledge concerning how these concepts are used in dialogues with clinical users.

Therefore, the whole problem has been cut into three separate sub-problems, each corresponding to a discipline and task:

1. **Clinical computational linguistics** – getting the language right
2. **Logical concept representation** – formal representation of concepts in ways which give rise to correct identification, classification, and retrieval of information in formal (computer) systems.
3. **Clinical pragmatics** – organizing information in ways expected by healthcare professionals and in ways that facilitate their daily work.

Because each of the three disciplines has developed separately and each uses different tools and techniques that are based on fundamentally different underlying principles, integrating such systems directly is difficult, perhaps impossible [18, 38]. The hypothesis of separability seriously underestimated the difficulties of integrating the three aspects of research outcomes in actual clinical information systems. One of the major difficulties in medical terminology has been the confusion of concepts and the words used to express those concepts. In particular, there exists ambiguity of mapping between the linguistic expressions and concept components in which one linguistic expression can be interpreted as more than one internal concept. Incorporating such ambiguity into formalisms and ontologies for clinical concept representation are hard.

VI. OVERVIEW OF THE FRAMEWORK

In order to address the deficiencies of the existing approaches semantic heterogeneity of medical information...
in a coordinated care enterprise environment, we proposed a framework, called Context-Mediated Semantics Interoperability (CMSI framework). The idea behind our framework is as follows:

1. Contexts should be explicitly represented; contextual knowledge should be associated with context representations; and contextual knowledge should guide all facets of an agent’s behavior.

2. A theory of medical data semantics should include multiple ontologies ranging from top-level generic ontology to application specific ontologies. Each ontology is associated with a context that ‘wraps’ around it. Semantics of a specific ontology is local to its context. Ontologies are related through the generalization/specialization relationship of their contexts, as well as through explicit ‘lifting’ rules.

3. With contextualization of data semantics, it is no longer required for two communicating agents (or data sources) to have common ontological commitment. Instead, we rely on context alignment and shared contextual knowledge to constrain semantic interpretation.

4. Contextualization hides the heterogeneity of data at the ontology level, just like ontologies effectively hide the heterogeneity of data at the syntax level.

5. Contexts and ontologies are two semantic coordination mechanisms for interoperability, with contexts taking priority over ontologies. In other words, commonality in contexts can over-ride heterogeneity in ontologies, but not vice versa.

Our CMSI framework is the integration and extension of the work on contextual ontology (C-OWL) by Bouquet [39] and the work on context schema (C-schema) by Turner [40]. It consists of the following six components:

- $\mathcal{C}$: a context space which is a set of contexts $\{C_i | i=1,..,N\}$, where $N$ is the total number of contexts
- $\mathcal{O}$: an ontology space which is a family of ontologies $\{O_i | i=1,..,N\}$
- $\Phi$: a set of inter-ontology bridging rules $\{\Phi_{ij} | (i, j \in \{1,.., N\})\}$.
  Each $\Phi_{ij}$ is a set of rules that specify how elements of ontology $O_i$ relates to elements in ontology $O_j$ if any relationship exists.
- $\Psi$: a set of inter-context bridging rules $\{\Psi_{ij} | (i, j \in \{1,.., N\})\}$ that specify how context $C_i$ relates to context $C_j$ if any relationship exists.
- $\Theta$: a set of rules governing context coordination.
- $\Theta'$: a set of rules governing ontology coordination.

Each of these components is again a complicated structure. For more details, see [35].

VII. POTENTIAL BENEFITS OF THE FRAMEWORK

The fundamental advantage of the framework originates from its premise that linguistic, conceptual and pragmatic issues of medical information must be considered together due to their close coupling in clinical interpretation and reasoning. In other words, we do not make any assumptions of separability as did in previous work (see section V). Instead, we make medical pragmatics explicit and use them as formal semantic coordination mechanisms that embrace semantic heterogeneity across different use scenarios and clinical practice specialties. This perspective can be stated into two assumptions below:

1. First, data are always produced with a given purpose. Medical data complexity and specificity is directly tailored to that purpose [41]. Medical information is entangled with its context of production in that the meaning, hardness and significance of a piece of information cannot be detached from the specific purpose that structured the gathering of that information.

2. Objects in medical data mutually elaborate each other, rather than being isolated “atoms.” For example, in the course of a patient’s illness trajectory, data items are constantly reinterpreted and reconstructed.

Modeling and representing context can lead to several benefits [42, 43]:

- **Economy of representation:** contexts can act as a focusing mechanism
- **Economy of reasoning:** reasoning can be performed with the context associated with an information source (instead of the whole data)
- **Managing inconsistent information:** As long as information is consistent within the context of the query of the user, inconsistency in information from different databases may be allowed
- **Flexible semantics:** An important consequence of associating abstractions or mappings with context is that the same two objects can be related to each other differently in two different contexts. Two objects might be semantically closer to each other in one context as compared to the other.

Our solution bares some similarity with the OpenEHR initiative (http://www.openEHR.org) [5, 25, 44]. Our context schema can be viewed as an extension of Archetypes in OpenEHR specification. Archetypes are agreed models of clinical or other domain-specific concepts. From a technical point of view archetypes are formal specifications of clinical content. From a clinical point of view, archetypes serve an intuitive means to define and discuss and present clinical content [5]. It empowers the health professionals to define and alter the accurate knowledge and information they need in the granularity they need.

We summarize specific benefits of CMSI framework to semantic integration of EMR systems:

1. **Supporting multiple points of view, poly-hierarchies and multiple levels of granularities.** This is in contrast to traditional ontology and terminology approaches that have limited applicability to a single use, single granularity, a single point of view, a mono-hierarchy and often to a single area of medicine. Unfortunately, reusable medical records require that terminologies support multiple points of view, poly-hierarchies and multiple levels of granularities.

2. **Open-Endedness:** Not only is medicine big, it is open-ended. There are constantly new discoveries in medical knowledge. The framework is design to manage changes of language with respect to the underlying concepts, clinical practice, and the underlying system of concepts itself.
(3) **Minimize the difficulties of achieving consensus.** Achieving clinical consensus on existing terminologies has proved particularly difficult. Physicians disagree. Nurses disagree. Healthcare professionals disagree. Our approach supports flexible levels of consensus appropriate in each area and what areas can be left for local choice.

VII. CONCLUSIONS

Following the proposed CMSI framework, a number of research questions must be fully answered in order to derive workable solutions for medical information service integration.

**Question 1** What are the appropriate attributes that define a context schema? We will incorporate the findings in context models literature [22, 45-52] and evaluate them using the insight we have achieved in earlier sections of this report to make proper choices. One important direction we will take is the activity-based representation of contexts, which has been explored before [21, 53].

**Question 2** What are the appropriate structure and level of granularity of context schemas? This is a central question that has to be answered based on deep understanding of the practical use of medical records, analytical reasoning of clinicians, and the structure of medical knowledge. I hope to develop a series of field studies observing and interviewing physicians in their workplaces.

**Question 3** What are the reasoning capabilities across-contexts that are needed by medical record applications? Context reasoning is important for real-time mediation of semantic exchange and negotiating meanings when heterogeneity exists. We will incorporate modal logic reasoning approach [51, 54] to address the complexity of reasoning on contexts.

Answering above questions require close interaction with healthcare delivery organizations (hospitals and patient care professionals) and observation of their practice use. We are in the process of developing concrete case studies and use scenarios through collaborations with local hospitals. These will be used as design artifacts for creating context schemas and alignment mechanisms.

ACKNOWLEDGEMENT

Funding of this research is partially supported by a grant from the Center for Enterprise Architecture Research, Penn State University.

REFERENCES


