ON RESOLVING THE PROBLEM OF SEMANTIC HETEROGENEITY IN CLINICAL INFORMATION SYSTEMS

Ching-Song D. Wei
Department of Computer Information Systems
BMCC, City University of New York, New York, USA
(212) 220-8385
cswei1231@optonline.net

Simon J. Doong
Department of Information Management
China University of Technology, Taipei, Taiwan
sdoong@cute.edu.tw

Sam Y. Sung
Department of Computer Science
South Texas College, McAllen, TX, USA
(956) 872-3443
sysung@southtexascollege.edu

Peter A. Ng
Department of Computer Science
University of Texas-Pan American, Edinburg, TX, USA
(956) 381-2472
ngp@utpa.edu

ABSTRACT

The input and output data of a Clinical Information System (CIS) depend heavily on the use of a medical database and an intensive Graphic User Interface (GUI). Currently there are a number of healthcare information standards, such as HL7 (Health Level 7, 2006) and Medical Markup Language (Medical Markup Language, 2006) to promote data exchange. However, it requires that both the sender and receiver of the data use the same standard for achieving the data exchange goal. For the structured and unstructured data to be available and interoperable among healthcare organizations, the well-known problems of data heterogeneity with the distributed database systems have to be solved. In this paper, we begin to describe the structured and unstructured text data in a Clinical Information System (CIS), and then focus on the problem of semantic heterogeneity. We exploit the ontology to represent and translate the semantic of data. Instead of using a translator for every pair of organizations, a set of shared ontologies is employed among multiple healthcare organizations to solve the semantic heterogeneity. In the process of mediation, a successful ontology mapping between two organizations reconciles the heterogeneity between them. Due to the nature of Web Service, semantic aggregation of Web Service is required to compose a Web Service from fine granularity Web Services according to the structure and the semantics of the composite Web Service output parameters.

Keywords: Clinical information systems, data interoperability, ontology and semantic web services.
1. INTRODUCTION

A Clinical Information System (CIS) is usually implemented using a Windows-based client-and-server architecture. The development of this system relies heavily on the use of a medical database and GUI for data input and output. The data source consists of multiple GUI’s fields of various forms, which compose structured data and free-text medical reports on patients. These free text medical reports are mostly made of the complicated part of unstructured text data. Usually, the free-text medical reports include reports on patient intake, examination, and discharge, which could be an input from a keyboard, a handwriting device, a voice microphone, transcription service or others. The current standards of domain knowledge, such as the Unified Medical Language System (UMLS) Metathesaurus (Aronson, 2001; Huang et al., 2003; Copper and Miller, 1998; Nakdarni et al., 2001) and the controlled vocabulary of Medical Subject Heading (MeSH) developed at the National Library Medicine (NLM), are some of the resources that could be used for unified medical terms and tools.

1.1 Dealing with Clinical Free Text

Much research is done in developing methods that can automatically map a clinical free text to concepts of standardized medical knowledge source such as UMLS Metathesaurus. Most of these systems employ various methods based on string matching, linguistic processing and utilization of part of speech tagging followed by identification of noun phrases. The Metathesaurus also provides a method which can be used to index text or to map terms from one vocabulary to another. It then discovers a set of UMLS concepts in a body of the index text and provides the ranking on these concepts in the set. Some of medical language extraction and encoding systems, such as MedLEE (Friedman et al., 2004), can be used to translate the medical text, written in a free-text, unstructured format, into a text of a standard and structured format.

Unstructured data source includes free text containing non-normalized biomedical terms, and the concept of the entire text should be extracted and mapped to a standard biomedical knowledge source such as UMLS. This is due to the fact that medical terms used by different healthcare providers are not unified. The Clinical Document Architecture (CDA) model (Dolin et al., 2001), which is a major effort for addressing interoperability, has established a common well-defined representational model for documents in different type of report and also for the concepts in the documents.

1.2 Inability of Relational Database in Medical Applications

Most of the CIS’s provide users with versatile functionalities such as creation, updating, and deletion of data records of a medical database, which is normally a relational database (RDB), to support the data manipulation. A RDB requires data to be specified in tabular form and conformed to a predefined schema and constraints. This data requirement promotes data integrity but discourages any changes on irregular data or data that evolves rapidly. The Relational Database Management System (RDBMS) will continue to be a dominant information management system, which manages all the critical enterprise data, although, it is difficult to support data sharing and data exchange in large-scale databases.

Furthermore, in traditional CIS, the GUI plays an important role for input and output between systems and their users. All the input data are captured by a GUI and then uploaded to the remote database server. Since each CIS is usually designed for a specific purpose and there is no unified term and structure among different CIS databases, it is very challenging to enable a system to interact with other systems of the same sort (such as, for the same purpose, area of application, etc.). It is a much more difficult task to provide interactions among different CIS systems, even within the same healthcare organization.

1.3 Data Sharing and Exchange among Medical Database Systems
The ability of sharing and exchanging data between database systems is a critical problem currently faced by many large healthcare organizations. A case study of supporting interoperability of autonomous hospital databases was presented by Zisman and Kramer (1997). It proposed a federation system which uses an unbalanced hierarchical structure to locate data in any federation within the system. But, this requires human assistance to define the terms that compose the hierarchy of terms, based on the user interests and applications, and to create meta-information about a database (schemas and instances) that could be shared with other federations within the system. Manolescu et al. (2001) proposed an architecture for integrating heterogeneous data source under an eXtensible Markup Language (XML) global schema (Bray et al., 2000). Then, XQuery is used to specify data retrieval and translated to SQL for each local data resource (XQuery, 2007).

The use of a standard-compliant XML for storing the structured and unstructured data on a GUI form provides a way for data sharing and therefore makes the data interoperability possible (Seligman and Rosenthal, 2001). With GUI, the use of XML as a basis for storing data in storage and for supporting data integration is continuously gaining acceptance in the medical research community (Ostell, 2005). The GUI form is similar to document data. The semi-structured data model of XML offers an appropriate way for storing the data. In order to maintain and obtain semantic interoperability, each value control must associate with its context information (Sciore et al. and Cheng et al., 1994). The context information is a metadata specifying the meanings, properties (such as quality, precision and source) and organization of its associated value of a value control. The use of XML for specifying the mapping description file eliminates the need of a specialized parser and a specialized description language and, therefore, simplifies the potential evolutions to the mapping specification language (Karadimas et al., 2000). In Wei et al. (2006), the extended semantic value model, in which each data value is contained in a value control, is associated with a set of attributes and a context. Then, following the semantics association path, a semantics association path, is used to create complete context information. The semantic association trees, which are of predefined semantic tree structures, describe the framework of property and semantic relations. Instead of storing it into a database, the content of GUI forms and controls, called data source, are exported as a file of a semi-structured XML format. The exported XML file contains only denormalized data and is ready and easier to be shared with other systems.

1.4 Semantic Web Service and Healthcare Data Interoperability

A Semantic Web Service based P2P interoperability infrastructure for healthcare information system was first proposed by Boniface and Wilken (2005) and then studied by Dogac (2006). It describes some of the difficulties for specifying Web Service standards and medical term standards which are key factors to make interoperability possible. In this system, the discovery and advertisement of services by healthcare organizations are managed through a peer-to-peer network structure; each health information system is represented by a node that communicates with a mediator or super-peer.

To achieve the goal of data sharing for the application in data interoperability between heterogeneous database systems, a clinical data repository can be stored, which is formed by applying a standard-compliant XML, with GUI form, for integrating these data. Currently there are a number of healthcare information standards, such as HL7 (Health Level 7, 2006; Dolin et al., 2001) and Medical Markup Language to promote data exchange. But it requires that both the sender and receiver of the data use the same standard in order to achieve the data exchange goal. In Wei et al. (2006), this coding method is adopted for transforming the free medical text into its associated code. The generic data XML model is developed and the feasibility for transforming the model into any of the standards is also demonstrated.

1.5 Data Heterogeneity Problem

Traditional direct-database-connection through a standardized interface posts many problems because of heterogeneous databases, heterogeneous attribute representations and semantics, and heterogeneous
schemas. For the structured and unstructured data to be available and interoperable among healthcare organizations, the well-known problem of data heterogeneity with the distributed database systems (e.g., Kim and Seo, 1991; Kashyap et al., 1996) has to be solved. This problem can be classified as: system heterogeneity, syntactic heterogeneity, structural heterogeneity and semantic heterogeneity. The system heterogeneity means that the applications and data may reside in different hardware and operating systems. The syntactic heterogeneity addresses that data source may use different representations and encodings. For achieving syntactic interoperability, compatible forms of encoding and access protocols could be used to allow information systems to communicate each other (Cardoso and Sheth, 2006). The structural heterogeneity states that different information systems may use different document layouts and formats, data structures, data models and schemas to store data. Finally, the semantic heterogeneity considers the intended semantic meaning associated with the content of an information item.

2. SEMANTIC WEB SERVICE AND HEALTH DATA INTEROPERABILITY

As discussed in Cardoso and Sheth (2006), XML supports the capability of solving the problem of syntactical heterogeneity: XPath (XPath, 2007) and XQuery provide ability to surpass some structural heterogeneity; and the recent standards of ontology representation languages, such as the Resource Description Framework (RDF) (Klyne et al., 2004) and the Ontology Web Language (OWL) (Smith et al. and McGuinness et al., 2004) provide an approach to solving semantic heterogeneity. To enable communications between two healthcare systems, a hard-coded method can be employed for writing all the required codes for translating messages between the two systems. But if a system is scaled up to compose a number of complex healthcare systems, the hard-coded approach would not be an efficient, useful and applicable solution. Instead, a shared ontology (Diaz and Baldo, 2005; Protégé, 2007) is used for solving the semantic heterogeneity problem. The shared ontology is a description and interpretation of semantic metadata that describes the relationships, rules and constraints of syntactic and structural metadata. In this paper, we investigate the application of ontology in translating the semantic of data among healthcare information systems, when Web Service technology is employed to overcome the system heterogeneity. The basic idea is described in Figure 2.1 in which a set of shared ontologies is used among multiple healthcare organizations, instead of using a translator for each pair of organizations. In the following section, we shall discuss the shared ontology.

![Figure 2.1: Using translators versus a shared ontology to achieve data interoperability](image)

2.1 Ontology

Ontologies are formal and explicit specifications of a shared conceptualization which consists of objects, concepts, and other entities in some areas of interest and their relationships among them (Gruber, 1993). The basic components in ontologies are concepts, relationships and individuals. The concepts are represented as sets that contain individuals, which meet the requirements’ specification for the membership of the concepts. The ontology concept has the same meaning as the class in object-oriented terms. For example, the concept Patient contains all the individuals which are patients in a given clinic domain. Concepts can be organized to form a superclass-class (or class-subclass) hierarchy, which is also known as taxonomy. Individuals represent objects (also known as instances) in a given domain. The concept’s individuals are equivalent to the class instances in object-oriented terms. Relationships characterize binary relations on individuals which link two individuals together. Relationships can have inverses. For example, the inverse of hasSibling is isSiblingOf. There are three types of relationships: object relationships, data-type relationships, and annotation relationships. An object relationship links an individual to another individual. A data-type relationship links an individual to a XML schema data-type value or a RDF literal. Annotation relationships are used to add information to classes of concepts. Relationships, which are represented as
links between any two individuals, may have a domain and a range. It means that the relationships link the domain’s individuals to the range’s individuals. As an example, Figure 2.2 describes a patient, John Smith, who is 25 years old, weighs 150 pounds, and has a brother whose name is Joe Smith. John Smith has disease pneumonia, and is a patient of Dr. John Dave. As shown in this figure, the relationships can be limited to having only single value but they can be either transitive or symmetric.

![Figure 2.2: A patient, John Smith with various relationships in an ontology](image)

The Protégé 3.2 OWL (Protégé, 2007) can be used to support the ontology development and maintenance for healthcare (the terms healthcare and clinical are used interchangeably) information that allows information sharing, exchange and interoperability among healthcare organizations. A logical model is employed that allows, in any definitions, building complex concepts from simple concepts.

### 2.2 Semantic Web Service

Web Services are modular, self-describing, and self-contained applications that are accessible over the internet. Besides supporting distributed computing development, they are considered to be a suitable solution for integrating disparate healthcare/clinical systems in a system heterogeneity environment. The key technologies used by Web Services are XML, Simple Object Access Protocol (SOAP), Hyper Text Transfer Protocol (HTTP), Web Service Description Language (WSDL) (Chinnici et al., 2003) and Universal Description, Discovery and Integration (UDDI) (UDDI, 2005). XML is one of the most popular standards for a platform that supports neutral interchange of structured data. SOAP is an XML-based messaging technology that specifies all the necessary rules for communicating with Web Services. HTTP is the most common protocol used by Web Services to transfer the SOAP-formatted documents between cooperating computers. A Web Service client can perform service discovery for finding out whether a Web Service exists and determining a way to interact with it. A Web Service uses WSDL to describe its operational information. But, WSDL does not contain semantic description of its operations. It only specifies the structure of message components using XML schema constructs. In our healthcare applications, the data exchange occurs only under the mutual agreement between two healthcare systems in different organizations or within the same organization. The rules of the legitimate policy of the Health Insurance Portability and Accountability Act (HIPAA) are enforced (HIPPA, 2007). The HIPAA’s security rules protect all clinical records and other patients’ identifiable health information maintained by any healthcare organizations.

One of the solutions for creating Semantic Web Services (Battle et al., 2005) is finding a way for mapping concepts in a Web Service description to ontology concepts. A user can explicitly define the semantics of a Web Service for a given domain. This approach provides a method for annotating the healthcare domain. When a Semantic Web Service is applied to perform the data exchange, the machine must be able to meditate the operations, input, and output of the target Web Service and to transform the data to a correct format and structure semantically and syntactically. We shall discuss the semantic annotation for Web Service in the following sub-sections.

#### 2.2.1 Semantic Annotation Architecture for Source Data
The process of associating semantic metadata with resources is called semantic annotation. A generic architecture of semantic annotations is described in Figure 2.3 (Nagarajan et al., 2006). Generally, the common resources could be Web page, text document, Web Service, or database. The semantic annotation of resources is supported by an existing domain model (ontology schema) and a knowledge base (ontology instances). Regardless of different resource types, the basic annotation process consists of three primary steps: entity identification, entity disambiguation, and annotation. The entity identification process involves extracting useful information from a document by applying various techniques of rule-based grammar, natural language processing, or user-defined templates or wrappers. The entity disambiguation is mainly to identify the correct references and representations of ontology instances. This process also considers various representations for the same entity, incorrect spellings, use of abbreviations, different name conventions, and so forth. Because our application includes the integration of database and ontology, it is crucial to have a well-defined method of entity disambiguation module (Blume, 2005). When entities are identified and disambiguated, the next step is the association of the entities in the resource with semantic metadata by applying the annotation process. The typical annotation languages include RDF and OWL.

### 2.2.2 Semantic Annotation for Web Services

In this paper, since the Semantic Web is chosen as the approach to enhancing data interoperability, we shall consider only the annotation of Semantic Web. The Semantic Web annotation process is different from resource annotation. In the domain of distributed healthcare/clinical information systems, the capacity of a code component anywhere on a network is described by WSDL. The WSDL describes a service as a collection of operational interfaces and their data type specification, and deployment information. For example, the patient demographic information can be represented by a schema in hospital A and a different schema in hospital B, as shown in Table 2.1 and Table 2.2, respectively.

<table>
<thead>
<tr>
<th>Field Name</th>
<th>MRN</th>
<th>LASTNAME</th>
<th>FIRSTNAME</th>
<th>MI</th>
<th>SEX</th>
<th>DATEOFBIRTH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date Type</td>
<td>int</td>
<td>string</td>
<td>string</td>
<td>string</td>
<td>string</td>
<td>date/time</td>
</tr>
</tbody>
</table>

Table 2.1: A Patient_Demographic schema

<table>
<thead>
<tr>
<th>Field Name</th>
<th>MRN</th>
<th>LNAME</th>
<th>FNAME</th>
<th>MIDDLE</th>
<th>GENDER</th>
<th>DOB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date Type</td>
<td>string</td>
<td>String</td>
<td>string</td>
<td>string</td>
<td>date/time</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.2: A Demographic_Description schema

Based upon the Patient_Demographic schema and the Demographics_Description schema, two WSDL documents can be generated for these two Web Services (Wei et al., 2008). Both systems do not have sufficient information for mapping from one schema to another schema and matching the semantic
meaning of their elements. Even though the WSDL document is of a widely accepted standard, the lack of a formal description of the meanings of their functionality and data of Web Services hampers the realization of integration promises. Additional work is required for eliminating the ambiguity of their interpretations. The application of annotating Web Service is to enable disambiguation, and automated service discovery and composition.

Basically, associating semantics with Web Service elements is to find the most appropriate semantic concept in ontology for a WSDL element. It can be done by matching WSDL’s elements of a schema onto a domain model schema described by OWL. Both WSDL and OWL are not based on the same model, and the OWL schema has more expressive capability than XML used by WSDL documents. Thus, the process has two different tasks, namely matching and mapping (Nagarajan, 2006). Matching between a WSDL’s XML and an OWL schema refers to as the process of finding the best semantic correspondences between elements of two schemas. Mapping addresses the actual representations of found matches and the transformation rules between elements of one schema to that of the other schema. For example, the weight information of a patient record of a hospital’s CIS which is sent to the other hospital’s CIS becomes meaningful provided that the exchanged data is accompanied with a process in terms of semantics of their functionality. It is common that the formats of the message exchange between these two CIS’s are incompatible. An example is the weight of a patient, one being in “pounds” and the other in “kilograms”. Thus, a mapping between these two elements of the schemas is required. This mapping specifies a transformation rule for converting the message from one format to another message format, i.e. from weight in pounds to an equivalent weight in kilograms because of the heterogeneities and conflicts. Possible heterogeneities and conflicts between input and output messages could make the generation of mappings between messages more difficult (Akkiraju, 2005).

Let consider creating mapping between the Web Service element and the domain model (or ontology concept) in which the Web Service elements are semantically associated. Based on this idea of using the domain model as a vehicle for inter-service communication, a generic architecture is given in Figure 2.4, that enables the interoperation between two CIS’s of two healthcare organizations.

![Figure 2.4: Using domain model (Ontology) for message transformation between two Web Services](image)

**2.2.3 WSDL Annotation**

In sequel, we shall discuss matching and mapping between a WSDL and OWL schemas and transformation between two ontologies.

WSDL does not offer any description for domain specific semantics such as service functionality semantics and service message semantics. This limits the usage of Web Services. Web Service Semantics - WSDL-S augments the expressivity of WSDL with semantics by employing concepts analogous to those in OWL-S (Martin, 2004) while being agnostic to the semantic representation language (Akkiraju et al., 2005). It provides an extension on elements pointing to external domain models such as OWL or RDF ontology.
Functional ontology can be used to classify coarse-grained Web Services in a healthcare domain, and message ontology can be used to annotate finer granularity services for retrieving corresponding data. When searching for right Web Services, users can consult with the functional ontology first to find out services they are looking for by using the functional semantics of the services. In order to organize and categorize the ontology structure, the industrial standard in healthcare domain HL7 is used as a hierarchy structure for service functionality ontology. Generally, it groups the service functionalities into the following clusters: Patient Administration, Order Entry, Query, Financial Management, Observation Reporting, Master Files, Medical Records/Information Management, Scheduling, Patient Referral, and Patient Care (Dogac et al., 2006). This ontology could be used to search for the right Web Services. Usually, an application of a Web Service in healthcare domain is an access to, update on, or insertion of parts of an electronic data record, which comprise service parameters. The data types and formats of these service parameters are defined in some message ontologies, which are classified under a node of functionality ontology for a Web Service. If a system of an institution does not use any standard, the widely accepted HL7 Clinical Document Architecture (CDA) (Dolin et al., 2000 and 2001) should be adapted to develop its message ontologies. If two institutions use two different standards, then they will be two different message structures, which require structural and semantic mapping between the message components in order to automate their interoperations.

2.2.4 Ontology Mapping

The ontology mapping process MAFRA is described in Maedche, Motik, Silva and Volz (2002). In brief, the process includes lift and normalization (Visser et al., 1997), discovery of the similarities of elements, semantic bridging, and execution. The lift and normalization module deals with syntactical, structural, lexical and language heterogeneity. Both source ontology and target ontology should be normalized to be their corresponding ontology with their ontology lexical elements that are described using a same representation language, such as OWL or RDF using the lift approach.

The elements’ similarity discovery includes the similarity of terms, which is also called lexical similarity (Syeda-Mahmood et al., 2006), and property similarity, which is responsible to acquire the similarity between concepts based on their properties, either the attributes or the relations of concepts. For discovery of the similarities of elements between schemas, accompanied by propagating the similarities from top to bottom when top-level has a higher similarity, the bottom-up method is used to propagate the similarities of elements from the lower parts of taxonomy to higher concepts.

For representing the similarity in a formal way, MAFRA provides a meta-ontology called Semantic Bridge Ontology (SBO) (Maedche, Motik, Silva and Volz, 2002). There are two types of semantic bridges, namely, the concept bridge and the property bridge. The concept bridge is used to define the semantic equivalence between source and target concepts. The property bridge specifies the semantic relations between source and target properties, either relations or attributes of concepts.

At the execution time, the instances of source ontology can be transformed to target ontology instances by evaluating the predefined semantic bridge. A condition for executing a semantic mapping is specified at development time and then to be verified at execution time by checking the required data values. Once the condition of a related concept bridge is verified to be true, the instances of the target ontology are created. After all the instances of concept bridges are generated, property bridges are created for each of concept bridges and the properties of target instances are set according to them.

3. CONCLUSION

The Healthcare Information Exchange and Interoperability (HIEI) between providers (hospitals and medical group practices), independent laboratories, pharmacies, payers, and public health departments
has been assessed and recognized as a highly promising benefit (Walker, 2005). Interoperability is a basic requirement of ensuring that the widespread Electronic Medical Record (EMR) adoption could yield the social and economic benefits for us. Without interoperability, the EMR adoption will strengthen the information vaults that exist in today’s paper-based medical files, resulting in even greater proprietary control over health information and, with it, control over patients themselves (Brailer, 2005).

In this paper, we proposed that a set of shared ontologies is used with Web Service and discussed how this information can be understood in different systems by reconciling heterogeneous attributes and their semantics. The representation and mapping of semantics will be elaborated to develop a standard framework for the environment of healthcare. Since the healthcare organizations are not expected to use standard format, a mechanism is devised for organizations to describe their formats and for the mediator to generate semantic mapping. Currently, we are prototyping a healthcare organizations including CIS’s using shared ontology to study the efficiency and effectiveness of the interoperability of the systems within the organizations. In addition, other issues such as the security and privacy policy will also be investigated.

REFERENCES


Additional References Available Upon Request from the Authors.