

Ontology Based Metadata Management in Medical Domains

Quddus Chong, Anup Marwadi, Kaustubh Supekar and Yugyung Lee

School of Interdisciplinary Computing and Engineering

University of Missouri – Kansas City

5100 Rockhill Rd. Kansas City, MO 64110

{qkc509, akm7bb, kss2r6, leeyu}@umkc.edu

Corresponding author: Yugyung Lee (816)235-5932, (816)235-5159 (fax)

In medical research, one of major challenges is to archive, access, and analyse various heterogeneous databases containing patient information gathered from a large volume of data over a long period of time. The main objective of this paper is (1) to examine how the use of metadata and ontologies can support the management and integration of clinical research databases and (2) to present our ontology-based metadata management system developed for the Cardiovascular Research Department at MAHI. Using the concept of web services, our system is globally accessible. The system is currently deployed on the existing medical information system at Mid America Heart Institute (MAHI) and used for clinical research.

Keywords: Ontology, metadata, integration, medical databases, UMLS, XML

1. INTRODUCTION

In medical research, investigators tend to work independently or in clusters of research teams. Raw data collected in experiments or clinical trials is usually stored in some electronic form, to take advantage of the powerful processing capabilities of scientific computing. Often, there is a need to exchange valuable information between different researchers or research groups, for the purpose of independent analysis or the verification of experimental results. Increasingly, we are also seeing the emergence of distributed scientific processing, for instance through GRID-based architectures (Bramley *et al*, 2000). The Internet provides an important platform for this activity of medical information exchange to take place. However, there are still some difficulties to resolve before seamless interoperability and interchange can occur. The main cause for these limitations can be viewed as the result of the diversity of computing platforms and data storage environments. Different research groups rely on different technical infrastructure to carry out their work. In the best case, these differences allow researchers to choose the most appropriate solution for their particular study. However, this non-uniformity has also given rise to the problem of how to integrate heterogeneous research data sources.

One promising line of approach is based on the use of metadata and ontologies. Metadata is information that captures the characteristics of instance data from a data source (Vaduva and Vetterli, 2001) possibly including the format and structure of the populated instance data, its

Copyright© 2003, Australian Computer Society Inc. General permission to republish, but not for profit, all or part of this material is granted, provided that the JRPIT copyright notice is given and that reference is made to the publication, to its date of issue, and to the fact that reprinting privileges were granted by permission of the Australian Computer Society Inc.

Manuscript received: 21 September 2002

Communicating Editor: Associate Professor Jim Warren

organisation, and its underlying conceptual context. Metadata is used in locating information, interpreting information, and integrating/transforming data. An ontology is an explicit specification of the conceptualisation of a domain (Fensel, 2001). Ontologies let domain experts, system developers, and applications perform reasoning about domain-specific information content. Standardised information models (e.g. HL7 RIM) and electronic vocabularies (e.g. UMLS) can be part of ontology. Ontologies allow the development of knowledge-based applications. Benefits of using ontologies include: facilitate sharing between systems and reuse of knowledge, aid new knowledge acquisition and improve the verification and validation of knowledge-based systems.

At the Mid-America Heart Institute (MAHI, 2003), there is an on-going effort to integrate various databases containing patient information gathered over 20 years of clinical studies. The data sources range from relational databases to flat text-based files and proprietary legacy systems. Their clinical research relies on definitions for concepts/conditions, which form the basis of their medical research, provided by various medical organisations like American College of Cardiology (ACC, 2003), Society of Thoracic Surgeons (STS, 2003) and Unified Medical Language System (UMLS, 2003), etc. They are used to measure clinical outcomes, facilitate internal quality improvement activities, meet regulatory mandates, and provide and individually utilise national benchmarks of comparative data.

Our task was to examine how access to these data sources could be facilitated through a centralised repository and analysis system. To that end, we considered a metadata management architecture and designed an ontology to classify and describe the data sources available in a clinical research environment. This allowed easier mapping between internal representation and standardised external representation of data. Such a system would be of remarkable use to MAHI for both data management and search/retrieval of the various concepts and definitions. In order to facilitate the entire process, we have developed an ontology-based UMLS Integration Project (OUIP) system that excels in both concept retrieval and data management purposes. The system semantically integrates concepts retrieved from heterogeneous sources into the MAHI data repository.

2. PROBLEM STATEMENTS

In general, most clinical researchers do not think in terms of underlying database representations, and prefer to perform information retrieval and analysis using familiar concepts from the medical vocabulary. It is the job of the information access system to transform the user's requests, stated in the domain-specific medical terminology, into the specific database queries. This points to the need for a *metadata management component* to coordinate the translation, maintain mappings between terminology and data source representations, as well as to keep the mapping information valid, as the underlying data source may change or get updated over time.

Information integration is the task of providing uniform and transparent access to the data managed by multiple databases (Lee, Bressan, Goh and Ramakrishnan, 1999; Calavense, *et al*, 1998). Broadly speaking, there are three types of data integration facilitated by the metadata management component, namely physical integration, logical integration, and semantic integration. *Physical integration* is the task of converting heterogeneous data sources into a common data format (such as XML) (Cluet *et al*, 2001). *Logical integration* is the activity of relating all data to common process models (e.g. HL7 model for a medical service like 'diagnose patient' or 'report outcomes'). *Semantic integration* allows cross-referencing, and sometimes inferencing, of data with regards to a common vocabulary or ontology (e.g. UMLS). In this work, we are interested in the logical and semantic aspects of integrating data from different databases under a common process model for clinical research.

The requirements for metadata in a clinical research environment include:

- Describe content-independent information (e.g. location of patient medical record), domain independent information (e.g. structure or format of the electronic patient record), and domain-specific information (e.g. guidelines for treatment of patients with acute myocardial infarction). The level of abstraction at which the information is represented is also described.
- Describe mapping and relationship between source data attributes and terms from a domain-specific vocabulary (e.g. “the data type *SK* in an Observation field in a patient record has relation *codedValue* to the term *Streptokinase* in the institution’s standardised medical lexicon”).
- Describe rules for interpreting information in specific context (“*Patients with both diabetes mellitus and coronary artery disease commonly have ischemic or hibernating myocardium without symptoms of angina*”).
- Capture/record/compute and store the information in a machine-readable format.
- Enable quick retrieval during processing via an index or hierarchical organisation structure.

The typical metadata management at medical database systems including MAHI has been manually performed: manually search for relevant concepts and type out the required ones into the data repository. This results in typographical errors, redundant entries and inconsistent values in the data repository where these definitions were stored. Such errors cause inconsistencies and incompleteness in medical analyses reports. Thus, it is highly critical to collect such medical concepts and definitions from disparate sources and then store them into a format according to local requirements. Here are the requirements, which we should take into consideration. The system should be flexible enough to detect already present concepts to avoid adding redundant information. It should also allow the user to add multiple definitions for the same medical concept i.e. semantically integrate concepts from different sources, so that either of them can be used for research. The system should be designed such that it does not need to modify the existing data format used at MAHI, rather it should convert the definitions into a format that is housed locally at MAHI.

3. RELATED WORK

Over the past several years, the field that has been most active in building and using ontologies (Gruber, 1993) has been Medical Informatics. There exist a large number of terminologies developed for different purposes (literature indexing and retrieval, electronic patient records, statistical reports on mortality, billing), in different subdomains (diseases, micro-organisms, diagnoses, medical devices, procedures, drugs). These terminologies have been built by different institutions (World Health Organisation, National Library of Medicine, College of American Pathologists, etc.). Each terminology has its own representation of the world, suitable for the purpose it has been developed for.

Ontologies are seen as a key component of terminologies interoperation, integration and information brokering systems (Amann *et al*, 2002). In van Heijst *et al*, (1997) knowledge engineering was described to construct ontologies as schematic descriptions of the contents of domain knowledge. Van Zyl and Corbett (2000) introduced a framework with support to compare multiple ontologies and their tools within different kinds of applications. Integration and reuse of ontologies can be evaluated using the framework for various applications including medical information systems. The *Unified Medical Language System* (UMLS, 2003) is one example of on-going efforts to define a comprehensive standardised electronic medical vocabulary for information sharing. The knowledge sources in UMLS have been used by a wide variety of applications programs to overcome retrieval problems caused by differences in terminology and the scattering of relevant information across many databases.

Recent metadata research focuses on data integration by understanding the schemas of the underlying data sources using domain ontologies. The metadata-based approach is relatively new to medical informatics systems. Recent trends in medical information integration and exchange have focused on federated database systems (e.g. the *Synapses* (Grimson, 2001)), or on the use of a data warehousing solution (e.g. the *CATCH* (Berndt *et al*, 2001) methodology). However, these approaches are dependent on a fixed record format or data warehouse schema to transform the source data into a global schema. The main disadvantage is that it is not suitable for frequent dynamic changes of schemas since the process of integration has to be repeated if a schema changes. Also, autonomy is sacrificed to resolve schematic conflicts, because databases have to reveal all information in their conceptual schemas, or alter its schema, to ease integration. Using a fixed global schema also restricts the user's ability to formulate queries based on different domain-specific semantics (Halevy *et al*, 2003). Other metadata-based database integration projects include *DataFoundry* (Critchlow *et al*, 1998), *DOME* (Cui *et al*, 2001), and *MIDAS* (Kayshap and Sheth, 2001).

In the area of scientific research, data is exchanged between organisations to collect raw data sets for testing and analysis. To support interoperability and provide better access, several metadata standardisation projects have been initiated. The Clinical Data Interchange Standards Consortium (CDISC, 2003) standard aims to develop an XML-based metadata model to support standard data interchange between medical and biopharmaceutical domains. Health-Level 7 (HL7, 2003) standard represents an effort to define an Electronic Patient Record (EPR) standard for the healthcare industry. EPRs represent patients' care lifecycle, ranging from epidemiology reports to insurance and billing claims and are also seen as the central component for Clinical Data Warehousing (Pedersen and Jensen, 1998). Hence, integrating data from different EPR systems is seen an important challenge.

Our approach is similar to these efforts by assigning a metadata component the task of defining the relationship between a domain ontology and the schemas of the underlying data sources. Our main contribution is chiefly in the application domain, namely applying a metadata-based architecture to databases and terminology used in cardiovascular clinical research.

4. OUR APPROACH

Now we describe our solution using an ontology-based resource brokering architecture applied to clinical research needs.

4.1. Towards Semantic Understanding between Knowledge Bases

Information management is becoming increasingly crucial in the medical and healthcare domain. This can be accounted for by: (1) the increasing use of Electronic Patient Records (EPR) to integrate the workflow of patient-related document management, (2) efforts to develop standards for data from diverse healthcare domains for the purpose of medical messaging or indexing, and (3) the increase in use of medical tools and diagnostic systems that record and store data of clinical interest electronically, for example on relational DBMS.

One method that has been proposed to manage the heterogeneity of medical data has been the use of ontologies, i.e. specifications of the conceptualisation of a domain, or conceptual schemas. Ontologies carry the benefit of allowing various entities represented in different data sources to have the same shared semantic meaning, taken from some domain of interest. Thus, queries for integrated knowledge across the different data sources become possible by posing the queries against the shared ontology.

It is expected that to be practical and not overly complex, ontologies will be developed by communities of users/practitioners who will not necessarily be knowledge engineers or logicians themselves. Data sources will then begin to resemble knowledge bases, allowing machine-understandable access through their semantic-based schemas. However, this creates the possibility of overlap again, as different ontologies may have equivalent concepts, or may in fact contain subsets of separate ontologies within themselves. To support data integration, and subsequent query operation, the separate ontologies must be classified and re-organised in a logical and semantic sense. Thus, the concept of metadata becomes helpful for managing data sources as well as their underlying ontologies. This points to a need for a formal model for metadata management of ontology-based data sources.

4.2. OUIP Data Elements

The data sources are discovered and shared through data registries that can be used by the medical community. Medical registries differ in different organisations. Some organisations specifically concentrate on cardiovascular data, whereas some concentrate on patient data etc. In order to perform quality research that can be published and accepted worldwide, medical organisations need to use data sources that are accepted and recommended by the international community.

In Table 1, we identify the types of metadata elements at the database and domain concepts levels. Here, we view a *resource* as being a conceptual mapping to a set of related entities. Resources are uniquely identified by a *resource identifier*. The role of the metadata manager is to maintain and ensure the validity of the mappings between named resources from the two layers. At each layer, the *resource representation* captures the current state of the resource. Retrievals are performed using the associated *representation interface*. The *resource model* describes the structure of the information being supplied, and is an important component for allowing interoperability. Finally, the *control data* for a resource indicates how the information requests should be handled based on the relationships within elements of a resource.

In the OUIP system, using standardised medical terms and concepts are maintained for more effective information sharing and exchange with external organisations while incorporating such accepted definitions into the codes are highly beneficial. The following are the resources currently used by the OUIP system at MAHI.

Data Element	DATABASE	DOMAIN VOCABULARY
Resource	Tables, objects, files, documents	Concepts/Terms from domain
Resource identifier	File name, key field	Ontology name
Representation types	Relational, Object-Oriented, Object-Relational, Flat file, XML	Frame logic (OKBC), Ontolingua, OCML, F-logic, LOOM, RDF(S), DAML+OIL
Representation interface	SQL, Xpath syntax, Vendor-specific	Description Logic language syntax
Resource model	Database schema	Top-level ontology
Control data	Referencing by key	Inferencing by is-a, part-of, connected-to, and other representation-defined relationships.

Table 1: Data Elements of a Metadata Management System

MAHI Database: MAHI uses the code-based approach to represent medical terms and concepts. The MAHI DR, a proprietary MAHI database, consists of codes used to represent information: *Code Table* houses the codes with their descriptions, types and the active status, assigned when a new concept integrated into the MAHI system. *Code Definition:* A code can have multiple definitions and thus library IDs are associated with them. This can easily allow the user to identify the source for a particular definition. *Code List Value:* A code might have list values (e.g., Weight might have list values 'Pounds', 'Kilograms', 'Grams', etc.). *Code Group Xref:* A code which might belong to a particular group (e.g., 'Patient Last Name', 'Patient Address' etc) also belongs to the 'Patient Demographics' group. Such references are made in this particular table. *Code Alias Table:* Codes may have aliases (e.g., Code 'Myocardial Infarction' has an alias 'MI'). *Code Library Table:* Currently MAHI uses concepts from the following libraries viz. the UMLS, the ACC and the STS. Each of these libraries are assigned unique library IDs in this table.

ACC Database: The American College of Cardiology (ACC, 2003) has provided a listing of 142 definitions, separated into various groups, as part of efforts to standardize Cardiovascular (CV) research and reporting. More than 27,000 cardiologists from the U.S. and around the world are members of ACC.

STS Database: The Society of Thoracic Surgeons (STS, 2003) has a classification of 242 definitions, subdivided into various categories. The aim of the STS definitions is to standardise information storage in cardiothoracic-related medical databases and, in fact, improve the quality and practice of thoracic surgery as a specialty.

UMLS Ontology: The UMLS Metathesaurus (2003) contains information about biomedical concepts and terms from many controlled vocabularies and classifications used in patient records, administrative health data, and others. Currently, the UMLS Metathesaurus includes 776940 concepts, 2.10 million concept names, and 11,137,725 relationships (source: UMLS Knowledge Sources, NLM, 2002). The Metathesaurus supports 26 separate data sources, including: ICD2002 ICD-90-CM, HCPT02-HCPCS Version of Current Procedural Terminology, and MSH2002 Medical Subject Headings. The Semantic Network contains information about the types or categories (e.g., "Disease or Syndrome," "Virus") to which all Metathesaurus concepts have been assigned and the permissible relationships among these types (e.g., "Virus" causes "Disease or Syndrome").

We constructed a customised ontology, called the OUIP ontology, to describe the meta-structure of the code information that is stored in the MAHI Data Repository. The OUIP ontology represents explicitly a shared understanding of the important concepts in the MAHI cardiovascular domain. They describe a conceptualisation of the cardiovascular domain in a knowledge representation formalism that can be used and shared among different health organisations (MAHI, 2003). The OUIP ontology serves as a cache for all the MAHI related concepts. It provides a mapping of a MAHI related concept definition to corresponding UMLS, STS and ACC definitions. In a way the local ontology serves as a customised ontology that supports the retrievals of cardiovascular vocabularies and definition from the various sources. Thus, any query made by the user first queries the local ontology. The polling service notifies for any new concepts and the local ontology is updated accordingly. This increases the reliability of service (since the ontology is housed locally) and avoids a connection to the external sources for every request made. The OUIP ontology was created using Protégé-2000 (2003) and stored in RDF (2003) format and Ontology schema is stored in RDFS (2003).

4.3. OUIP Semantics of Metadata Integration

In this section, we describe the semantic model for metadata integration that is used in the MAHI OUIP system. We consider a layered approach to metadata management where the layered stack

consists of: (1) the resource layer, (2) the domain vocabulary layer, (3) the mapping layer, and (4) the context layer.

A resource is a module of informational content, and is the base physical representation of data in the metadata management framework. At the resource layer, resources are formally stored in the form $R = \{O, P, T\}$, where O is the object representation of resource R , P is the parent object of O , and T is the type or schema of the resource. This allows the creation of a tree-structure for representing data sources in terms of hierarchical parent-child relations between resources. Thus, for example, in the XML data source containing: `<patient><pid>1123</pid></patient>`, the resource representation for element `<pid>` has the Xpath (2003) expression `to <pid>1123</pid>` as the object, the Xpath expression to the enclosing node `<patient>` as parent, and the XML schema fragment for the `<pid>` element as the type of resource. Relational resources and Object-oriented resources are modeled intuitively as well using this tuple format.

A domain vocabulary is a set of statements S describing the concepts and relationships between those concepts in the ontology for the domain of interest. A statement in a domain vocabulary S is formally represented in the triple form $S = \{sub, prop, val\}$, where *sub* is a concept representing the subject of the statement, *prop* is a relationship (also a concept) representing a property of the subject, and *val* is a concept or a literal representing the property's value. Since *sub*, *prop*, and *val* can be defined with statements, the triple format is sufficient to describe all concepts and linking relationships between them, given that each statement is assigned a unique namespace as an identifier.

Mappings store how data source elements and domain vocabulary concepts are linked, as well as how domain vocabulary concepts are linked between different vocabularies. There are two main categories describing mappings, namely: (1) mappings from data source to domain vocabulary (resources-to-concepts), and (2) mappings from domain vocabulary to another domain vocabulary (concept-to-concept). For resources-to-concepts mapping, there are four cases to consider:

1. A single resource is used to instantiate a single concept, and the (semantically equivalent) concept is found in one or more vocabularies.
2. A resource is used to instantiate a collection of concepts, and the concepts are found in one or more vocabularies.
3. A set of resources are used to instantiate a single concept.
4. More than one resource, each representing the same semantically disjoint concept.

We formalise this category of mapping thus. A mapping of resource-to-concept is denoted $Mapping_{RC} = \{type, E_R/A_R, E_C/A_C\}$, where *type* indicates one of the above cases, E_R is a set of equivalent resources (one or more), A_R is a set of resources to be aggregated, E_C gives the set of equivalent concepts (one or more), and A_C represents a set of concepts to be aggregated. In the case of a single resource to single concept mapping, there should be only one resource in E_R and one concept (the one it instantiates) in E_C . We allow the listing of more than one resource in E_R and more than one concept in E_C in the same mapping, if the arities are the same. This is treated as an ordered list of one-to-one mappings.

The second category of mappings is more challenging, as it requires some method for ontology integration. One such technique has been proposed in Calvanese, Giacomo and Lenzerini (2001). Here we will consider this category of mapping to be best specified as additional (mapping) statements to be stored at the domain vocabulary layer.

The context layer stores the rules to determine if a mapping is applicable. This can be represented formally as context rule $CX = \{E, C, A\}$ where E is a set of one or more events to be monitored, C is the set of one or more condition(s) to check against when the context is triggered, and A is the set of one or more action to be performed if the condition(s) evaluate to true. We can

adopt the concept of Active Rules here. In an Active Rule, the context is represented as declarative statement:

ON some event e
IF some condition c
DO some action a

Thus, context is understood here to be a rule that is fired to perform some allowed actions if and only if a set of conditions holds true. The primary event type that will be monitored for is the Query operation. Note that Active Rules may be nested, such that an event could trigger subsequent rules to be activated. The conditions that cause the rules to be executed may include: (1) environmental conditions (e.g. the detection of whether a data source is currently active or not), (2) semantic conditions (e.g. the query source is from a medical application therefore only vocabularies from medical-related domains are considered unless otherwise stated), and (3) logical conditions (e.g. the existing data sources can only supply a partial result to the query, and thus the query has to be separated into one or more queries for the answerable result). Actions here are operations that generate a view of the global schema by activating the mappings that are permissible under the conditions evaluated earlier.

4.4. The OUIP Architecture

We have developed the Ontology-based UMLS Integration Project (OUIP) to provide transparent access to heterogeneous data sources including Patient Admit files, and databases for Cath Lab information and drug-specific studies. The metadata management component unifies and organises the data sources by maintaining structural and semantic information about each source, recording the relationships between attributes of the data sources with terms from a medical vocabulary, and computing contextual information gleaned from these linkages and other resource-related information. The contextual information is what allows the information retrieval system to generate queries across heterogeneous sources and return only the information matching to the researcher's request criteria. Figure 1 shows the design of the architecture.

The *Resource Broker* is responsible for determining what data sources should be contacted to meet a high-level user request for information. This is determined by searching through the Resource Directory for metadata about data sources that match the request criteria. The parameters are further refined based on constraints identified by the Contextual Knowledge Base. Then the relevant metadata is returned to the Query Processor, which sends the low-level request to the identified data sources.

The *Metadata Repository* handles the storage and management of the many types of metadata used by the Resource Broker. The Metadata Repository is contacted by the Resource Broker when a request for information is made by the user via the Query Processor. The Metadata Repository supplies the Resource Broker with the resource identifiers and resource attributes relevant to the query request, as well as any additional contextual information that pertains to retrieving and interpreting those resource elements under the request conditions specified by the user. The Metadata Repository also performs the polling service for Vocabulary Knowledge Base such as UMLS using scheduling mechanisms.

The *Resource Directory* supplies the actual description of each database's content. This information is stored in a common meta-model format. The use of a directory allows the attributes stored across different data sources to be indexed and allows faster resource look-up. Content

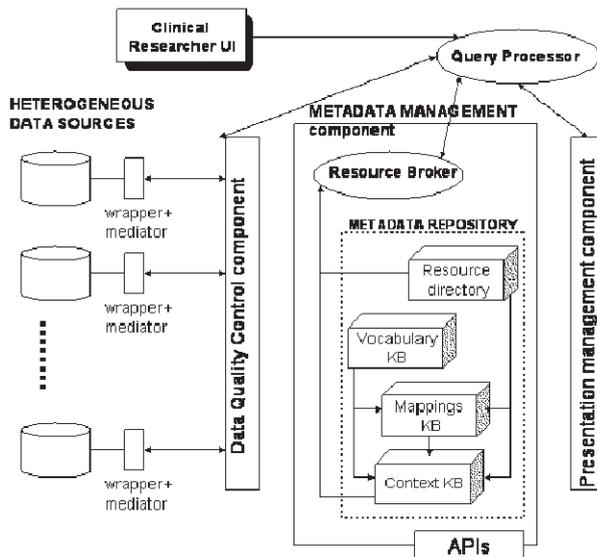


Figure 1: The OUIP System Architecture

descriptions here include parent-child relationships between attributes (for object-oriented or XML data sources) and table-column relationships (for relational data sources), a human-readable description of the attribute, and the syntax information to query that attribute (possibly a reference to a stored SQL procedure or Xpath expression). The directory also stores content-independent information such as the host server location for the data source.

The *Vocabulary Knowledge Base* supplies the domain-specific ontologies used to semantically define concepts and terminology related to the clinical researchers' work. This allows us to store different ontology models and use the one most relevant to the user's request for information. For instance a separate ontology for describing *Patients as Healthcare Customers (Patient Billing ontology)* or *Patients as Clinical Trial Participant (Clinical Trial ontology)* could be used to integrate heterogeneous data sources containing patient information, and generate appropriate contextual information. This is also useful for performing information focusing when retrieving information, because we can make use of different ontology models that describe the same domain concept at different levels of abstraction. For instance, a researcher investigating heart failure can specify whether the level of abstraction should be at the *Population, Individual, Organ, or Molecule* level. Finally, we can combine ontologies to search for concepts related to more than one domain model.

The *Mappings Knowledge Base* stores the mappings between the classes or properties in different ontologies, as well as correspondences between the attributes in the physical data sources with the abstract concepts represented in the ontologies. Generally one data source attribute can be related to properties in one or more ontologies. Currently, the administrator has to specify the mapping relationships, but inferencing based on data-mining techniques could be applied to automatically obtain mappings (Fowler and Martin, 1997).

The *Context Knowledge Base* stores information related to the how the data source attributes should be interpreted under specific request criteria. This knowledge base contains rules for what constraints to apply to the resource broker when it composes the metadata for the query processor.

Inferencing is based on the supplied request, mapping information, and relationship between terms in the relevant ontologies. This helps to refine the parameters for the data source queries, so that the level of semantic accuracy is improved. It also makes the system more adaptable, as new context rules can be added to handle different types of information requests.

We include well-defined *Application Programming Interfaces* (APIs) to the metadata management component to make the available metadata services (Resource Broker and Metadata Repository) accessible to other systems or applications. This allows the *OUIP* to be extensible, beyond the basic functionality of user-entered queries.

5. THE OUIP IMPLEMENTATION

5.1. Implementation Environment

The OUIP system runs upon the Windows NT/2000 Operating Systems. Specifically, the OUIP architecture consists of the following technologies: ASP.NET web forms (using C#) for Presentation layer, .NET web services, .NET components (using C#) for Middle/Business layer and SQL Server 2000 and XML (using ADO.NET access) for Data layer. The MAHI ontology is represented in RDF and Jena-1.4 tool is used for RDF query and update. Using the concept of web services, interface to the OUIP system was easily made through HTTP applications like web browsers via the use of XML messages thus providing a global access to the system. The OUIP system can be accessed globally at MAHI (2003).

5.2. Performing Queries on OUIP

The OUIP Query interface (Figure 2) allows the user to query for the definition of a medical term by submitting the query as a HTML form. The form components shown in the page includes: *Concept Text* defines user types in the term to obtain the definition (e.g. *Myocardial infarction, hypertension*); *Timeline* describes that definitions stored in the MAHI Data Repository are assigned

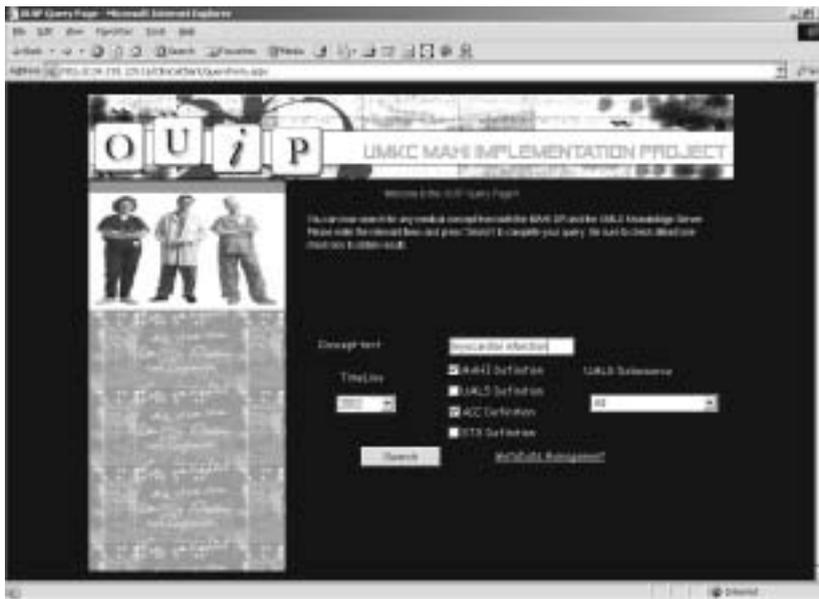


Figure 2: The OUIP Query Interface

a date when they are activated. Inactive definitions are still maintained in the Repository, with the date of inactivation. The Timeline drop down box allows the user to specify that only definitions active during the selected year should be retrieved; *Definition libraries* describes the user can select one or more libraries to search for a definition to a medical term. In the current implementation, the available libraries are: *MAHI* slot for definitions available from the MAHI Data Repository; *UMLS* for definitions available from the Unified Medical Language Server. The user can select a specific UMLS data source to query. The available data sources are: *Health Level 7 (HL7)* vocabulary, *ICD-9-CM*, *ICD-10*, *MEDLINE*, *MeSH* (Medical Subject Headings), *SNOMED* (Systemized Nomenclature of Medicine), *Neuromenes Brain Hierarchy*; *STS* slot for definitions available from Society of Thoracic Surgeons; *ACC* slot for definitions available from American College of Cardiology; The *Metadata Management* link leads to a form for adding, modifying, and deactivating code definitions (see Section 5.3 for more details).

If the concept entered by the user is discovered in one of the definition libraries, its definition is returned to the user on the OUIP Results page (Figure 3). In general, results are returned with the following information: The *Library definition ID* describes the unique identifier for this term according to the classification used by the definition’s source library.

In UMLS, the unique identifier is the CUID, while ACC and STS have their own IDs respectively; The *Library definition Name* is the name assigned to the concept. Different classifications may have different names for the same concepts (i.e. synonyms); The *Library Definition* is a description or explanation of the meaning of the term, its proper usage, and/or the process to which the term belongs; The *Library specific fields* describe additional information such that for instance, ACC and STS include a group name as part of the definition returned. Concepts from UMLS are linked in a semantic network. Every medical term can be associated with terms that satisfy a ‘broader than’ or ‘narrower than’ relationship. For instance, ‘*Acute myocardial infarction*’ is a narrower term or specialization for the medical concept ‘*Myocardial infarction*’. Similarly, ‘*Cardiovascular diseases*’ is a broader term or generalisation of ‘*Myocardial infarction*’. OUIP

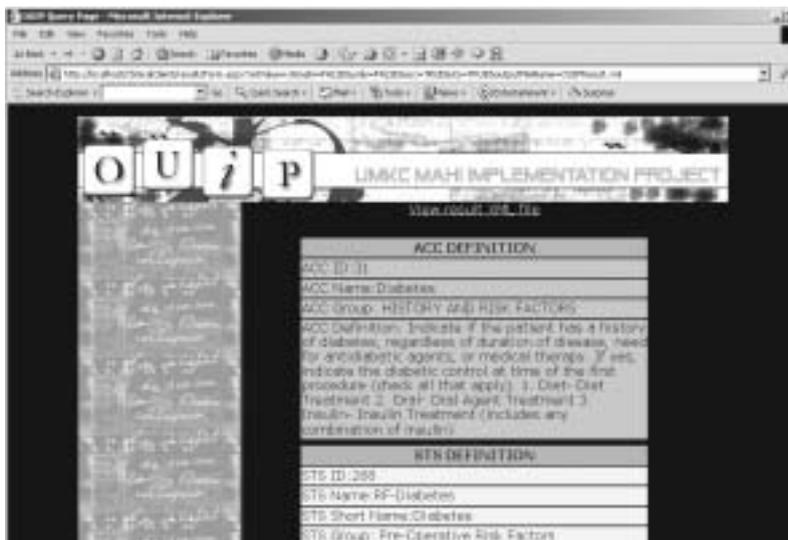


Figure 3: The OUIP Query Result Interface



Figure 4: The OUIP Query XML Result Interface

provides the user with a way to view these related terms. The results of a query may also be viewed in XML format as shown in Figure 4.

5.3. Metadata Management

The Metadata Management user interface (Figure 5) allows the user to add a new concept to the MAHI Data Repository, to add a new definition to an already existing concept, and to add list values

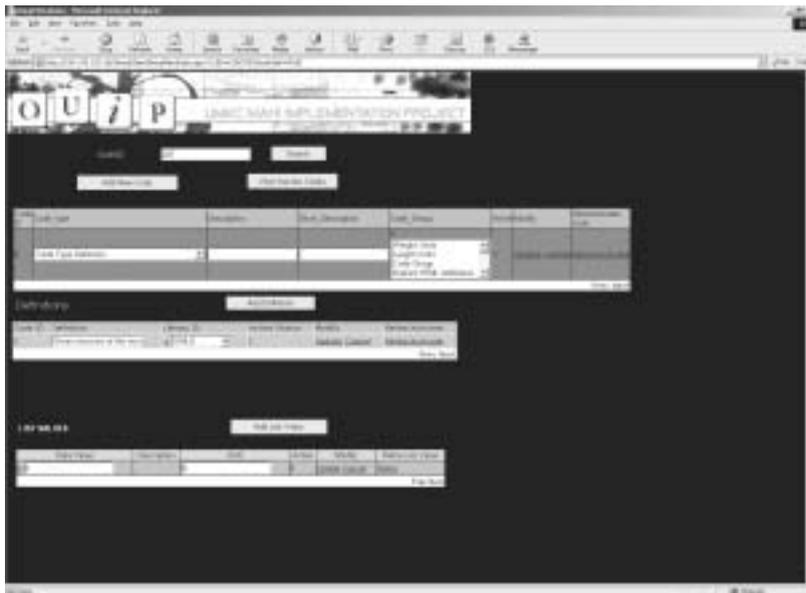


Figure 5: The OUIP Metadata Management Interface

to customise user interface forms. In some cases, a term being queried might not exist in the MAHI Data Repository. If a query returns a match from the UMLS, ACC, or STS libraries, the user may choose to incorporate this definition into the MAHI Data Repository. If a concept is already stored in the MAHI Data Repository, the user may choose to add supplemental definitions from the UMLS, ACC, or STS libraries.

Adding a new code is synonymous with adding a new concept to the MAHI Data Repository. Once the Code Entry form component is displayed, the user can enter in a new description and short description for a concept and assign it to a *MAHI Code Type*. The concept can also be associated with a *Code Group* that allows the concept or code to be used as a unique category for some grouping of values. If the code is currently active, it will be displayed with an *Active* status value of 1, or 0 if inactive. The user can also view a list of all currently inactive code, modify the code entry or cancel changes, and retire or activate a code entry. A list is a group of values that may appear as user options in a customised user interface. The metadata management interface allows the user to add a value to a predefined user interface list in the MAHI Data Repository.

5.4. Use of the OUIP System at MAHI

The OUIP system aims to provide a useful medical resource to physicians, researchers and biostatisticians at MAHI. Specifically, the system helps to provide a common user interface for retrieving MAHI code definitions in terms of their relation to existing standardised medical vocabularies. Currently, we have deployed the OUIP system at Mid America Heart Institute (MAHI). The major stakeholders of the OUIP system as depicted in Figure 6 are Medical Researcher, Knowledge Engineer, Database Administrator, and Clinical Application Systems. The following use cases from different stakeholders’ perspective strengthen our claim of extensive applicability of the OUIP system at Mid America Heart Institute.

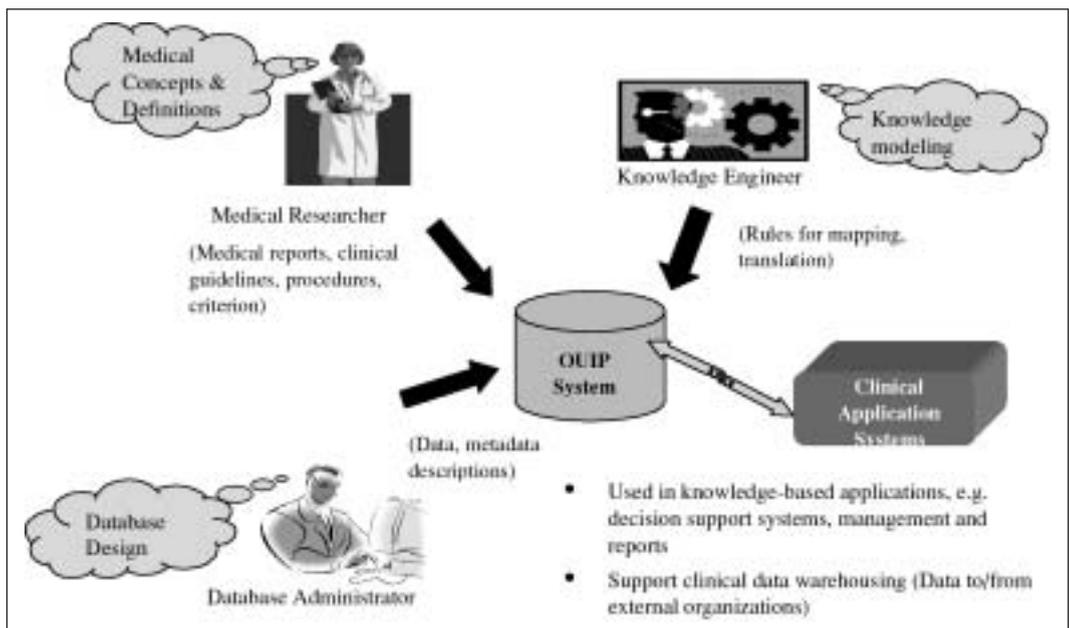


Figure 6: The OUIP usage at Mid America Heart Institute

Medical Researcher: Major stakeholders of the OUIP system are medical researchers. Consider a scenario. Dr. Johnson is preparing a paper about *acute myocardial infarction* for a cardiovascular journal publication. His paper is based on the report that the MAHI database administrator generates. However the report contains MAHI specific term “Acute MI” for a concept. He uses the OUIP system to retrieve corresponding ICD-10 code (International Classification of disease) for the MAHI specific concept “Acute MI” and describes MAHI specific concept "Acute MI" referred as ICD-10 code “I21”. Corresponding definitions can be further obtained from other sources such as MeSH or MEDLINE through the OUIP UI. This scenario demonstrates that the OUIP system allows him to retrieve information about cardiovascular terminologies via a customised mapping between internal representation and external representations.

Knowledge Engineer: The knowledge engineer (KE) is responsible for creating and maintaining the OUIP ontology. KE also performs Metadata management through the OUIP user interface. As part of metadata management KE specifies mapping rules between local schemas and local access operations into global schemas and access operations into another model. The metadata knowledge provides enough information regarding the semantics and source attributes. In addition, KE defines translation rules for data. For example, A term *Cardiovascular Disease* borrowed from UMLS data source has broader concept *DISEASES*. It means that the data has a source UMLS and needs to be translated and stored in the *DISEASES* hierarchy when storing it locally. The OUIP system allows him/her to add multiple definitions for the same medical concept i.e. semantically integrate concepts from different sources.

Database Administrator: The database administrator is mainly concerned with creating and updating database schema associated with the OUIP metadata. The OUIP system helps him/her to understand the semantics of underlying data elements and schemas.

Clinical Application Systems: Major sets of clinical applications systems that would benefit from the OUIP system are medical decision support systems, cross-organisational collaborative clinical research, and clinical data warehousing. For instance, Mid America Heart Institute is working with various medical communities across the United States as a part of Cardiovascular Outcomes Research Consortium (2003). The main objective of this consortium is to assist delivery of healthcare today and provide a translation infrastructure to eradicate disease in the future. The research focuses on facilitating seamless interoperability between various medical institutions that have different semantic interpretation of the same domain. The OUIP system would play the role of a language mediator for the communication amongst such semantically variegated medical institutions.

6. CONCLUSION

This paper presented an ontology-based metadata management system, OUIP, which integrates data from heterogeneous sources such as ACC, STS and UMLS. The ontology-based metadata management system allows a customised integration of heterogeneous clinical databases. The automatic mechanisms employed by the system to maintain the metadata of the MAHI databases were shown to be promising. The deployment of the OUIP system at the Cardiovascular Research department at MAHI (2003) verified the feasibility of the proposed OUIP model. Further, flexible presentation (a user-specified view) and data quality control (for data accuracy and integrity) are very critical issues in the clinical research environment. The OUIP system is being implemented to incorporate data quality control and presentation components, which would increase the accuracy of information management, into the OUIP system.

REFERENCES

- ACC (2003): The American College of Cardiology Website. <http://www.acc.org>. Accessed 30 January 2003.
- AMANN, B., BEERI, C., FUNDULAKI, I. and SCHOLL, M. (2002): Ontology-based integration of XML web resources. I. HORROCK and J. HENDLER (Eds): ISWC 2002, LNCS 2342, 117–131.
- BERNDT, D.J., FISHER, J.W., HEVNER, A.R. and STUDNICKI, J. (2001): Healthcare data warehousing and quality assurance. *IEEE Computer*, 34(12), 56–65.
- BRAMLEY, R., CHIU, K., DIWAN, S., GANNON, D., GOVINDARAJU, M., MUKHI, N., TEMKO B. and YECHURI, M. (2000): A component based services architecture for building distributed applications. In *Proc. 9th IEEE International Symposium on High Performance Distributed Computing*. Pittsburgh, PA, August.
- CALVANESE, D., GIACAMO, G.D. and LENZERINI, M. (2001): A framework for ontology integration. In *Proc. of 1st Semantic Web Working Symposium*, 303–316.
- CALVANESE, D., De GIACOMO, G., LENZERINI, M., NARDI, D. and ROSATI, R. (1998): Description logic framework for information integration. In *Proc. of the 6th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR'98)*, 2–13.
- CDSIC. (2003): Clinical data interchange standards consortium website. <http://www.cdsc.org/>. Accessed 30 January 2003.
- CLUET, S., VELTRI, P. and VODISLAV, D. (2001): Views in a large scale XML repository. In *Proceedings VLBD*, Rome, Italy, 271–280.
- CRITCHLOW, T., GANESH, M. and MUSICK, R. (1998): Meta-data based mediator generation. *Proceedings of Conference on Cooperative Information Systems*, 168–176.
- CROC (2003): Cardiovascular research outcomes consortium. <http://www.saint-lukes.org/mahi/html/general/ClinicalTrials/OutcomesConsortium.htm>. Accessed 30 January 2003.
- CUI, Z., JONES, D. and O'BRIEN, P. (2001): Issues in ontology-based information integration. *Proceedings of IJCAI-01 Workshop on E-Business & the Intelligent Web*.
- FENSEL, D. (2001): Ontologies: Silver bullet for knowledge management and Electronic Commerce. Springer-Verlag.
- FOWLER, J. and MARTIN, G. (1997): The healthcare administrator's associate: An experiment in distributed healthcare information systems. *Proceedings of AMIA Fall Symposium*.
- GRIMSON, J., STEPHENS, G. and JUNG, B. (2001): Sharing health-care records over the internet. *IEEE Internet Computing*, 5(4): 49–58.
- GRUBER, T.R. (1993): Toward principles for the design of ontologies used for knowledge sharing. In N. GUARINO and R. POLI, (eds.) *Proceedings of International Workshop on Formal Ontology*.
- HALEVY, A, IVES, Z.G., SUCIU and D. TATARINOV, I. (2003): Schema mediation in peer data management systems. To appear, *International Conference on Data Engineering*.
- HL7 (2003): The health level 7 website. <http://www.hl7.org>. Accessed 30 January 2003.
- KAYSHAP, V. and SHETH, A. (2000): Information brokering across heterogeneous digital data. Kluwer Academic Publishers.
- LEE, M., BRESSAN, S., GOH, C.H. and RAMAKRISHNAN R. (1999): Integration of disparate information sources: A short survey. *ACM Multimedia*, 155–158.
- MAHI (2003): The mid America heart institute website. <http://www.saint-lukes.org/mahihp.asp>. Accessed 30 January 2003.
- OUIP (2003): MAHI OUIP Server. <http://sice527.ddns.umkc.edu/ClinicalClient/index.html>. Accessed 30 January 2003.
- PEDERSEN, T.B. and JENSEN, C.S. (1998): Research issues in clinical data warehousing. In *Proceedings of the 10th International Conference on Scientific and Statistical Database Management*. 43–52. IEEE Computer Society.
- PROTÉGÉ (2003): The Protégé project website. <http://protege.stanford.edu/>. Accessed 30 January 2003.
- RDF (2003): The resource description framework website. <http://www.w3.org/rdf/>. Accessed 30 January 2003.
- RDFS (2003): The W3C RDF Schema website. <http://www.w3.org/TR/rdf-schema/>. Accessed 30 January 2003.
- STS (2003): The society of thoracic surgeons website. <http://www.sts.org>. Accessed 30 January 2003.
- UMLS (2003): The unified medical language system website. <http://www.nlm.nih.gov/research/umls/>. Accessed 30 January 2003.
- VADUVA, A. and VETTERLI, T. (2001): Metadata management for data warehousing: An overview. *IJCIS* 10(3): 273–298.
- XPATH (2003): The W3C XPath website. <http://www.w3.org/TR/xpath>. Accessed 30 January 2003.
- VAN HEIJST, G., SCHREIBER, A.T. and WIELINGA, B.J. (1997): Using explicit ontologies in KBS development, *International Journal of Human-Computer Studies*, 46, 183–292.
- VAN ZYL, J. and CORBETT, D. (2000): A framework for comparing methods for using or reusing multiple ontologies in an application, the proceedings of the *Conceptual Structures Conference of 2000: Working with Conceptual Structures*, Shaker-Verlag, Aachen.

BIOGRAPHICAL NOTES

Quddus Chong is a staff software engineer with IBM Data Management/Informix R&D group in Lenexa, Kansas. He is currently completing his MS in Computer Science at University of Missouri in Kansas City. His research projects have included community-based learning systems, metadata management, heterogeneous medical data integration using ontologies, autonomic grid instrumentation of web services, and Active Rule based transactional web workflows. (913)599-7103. (qchong@us.ibm.com)



Quddus Chong

Anup Marwadi received his B.Sc. degree in Computer Science from the University of Mumbai, India and received his M.Sc. degree in Computer Science from the University of Missouri, Kansas City, with a thesis in 'Ontological Semantic Integration Model'. His research interests include Medical Informatics, Ontologies, P2P and Web Service Oriented Architectures. He is currently working as a Solutions Architect at National Research Center for College and University Admissions where he designs and develops software solutions to promote student recruitment processes.



Anup Marwadi

Kaustubh Supekar received the B.Sc. degree in computer engineering from the University of Mumbai, India. He is pursuing Masters in Computer Science in School of Interdisciplinary Computing and Engineering at the University of Missouri, Kansas City. He is currently working at Distributed Intelligent Computing Lab and is doing research on Semantic Web, Web Service based workflow systems, P2P, Ontology searching and classification.



Kaustubh Supekar

Yugyung Lee is Assistant Professor at the University of Missouri at Kansas City. She received a B.Sc. degree in Computer Science from the University of Washington in 1990 and a Ph.D. degree in Computer and Information Sciences from the New Jersey Institute of Technology in 1997. Before joining the UMKC, she was working at MCC as a research scientist. Her research interests include Semantic Web and Knowledge Discovery, distributed systems (Middleware, Peer-to-Peer, Grid, etc), distributed data mining, agent-based computing and architectures, Web technologies, and Medical Informatics.



Yugyung Lee