Harnessing Knowledge from Ad Hoc Queries by Creating a Zone of Standardization

Jehad Alomari[1], Farrukh Mohammed[2] , and Kathleen Hayden[3]

[1]Department of Computer Science and Information System, Taif University, Taif, KSA
ahal8@yahoo.com
[2]University of Michigan Health System, Ann Arbor, Michigan, USA
farrukt@yahoo.com

[3]University of Michigan Health System, Ann Arbor, Michigan, USA
kathayd@med.umich.edu

ABSTRACT

Business Intelligence drives strategic business decisions by providing timely and accurate information. Beneath the demand for constantly changing information-needs usually lay a highly complex entangled mesh of business rules, data sets, query logic and custom-ad hoc Reports. The discrete nature of ad hoc reports may lead to inefficiencies in processing times and also valuable business knowledge is being lost. If the ad hoc queries are viewed within a context of larger collection of queries we could derive meaningful information worthy of reuse. Expert Query System (EQS) is an approach that drives ad hoc queries towards standardization, structure and documentation to reuse, recycle queries, harness knowledge and reduce the foot prints of the subsequent ad hoc queries. Therefore, a comprehensive, real time, manageable, and knowledge-based solution for ad hoc reporting is absolutely necessary. We proposed an expert query system based on a common metadata model that incrementally matures itself and presents information to the users in relative terms to empower business intelligence solution.

KEYWORDS

Database Management, Database Architecture, Expert System, Query Design and Implementation, Query Processing, and Query Formulation

1. INTRODUCTION

Capturing the relevant experience and knowledge is essential to sustain continual growth and to maintain a competitive advantage. Business Intelligence and Analytics tap into the knowledge to provide timely and accurate information, enabling the enterprise in decision making. Beneath the analytics is the demand for constantly changing information-needs which translates to a highly complex entangled mesh of business rules, data sets and query logic. Many technologies such as artificial intelligent systems and expert systems have been utilized to perform efficient and effective data processing on complex problems in various domains. In 1986, Peter Jackson defines an expert system as " a computing system capable of representing and reasoning about some knowledge–rich domain, such as internal medicine or geology, with a view to solving problems and giving advice" [2]. Expert systems (ES) are a subfield of applied artificial intelligence (AI) that incorporates knowledge and analytical skills from domain experts to analyze information about a specifc problem. ES is knowledge bases software that provides expertise oriented advices and explain the logic behind them to solve a problem [3, 6, 8]. Solving a problem by relying on manual data analysis to turn data into knowledge is becoming unfeasible because data volumes grow exponentially [4, 5]. Therefore, processing a large volume of data requires advanced computer tools capable of adapting to the user requirements

DOI : 10.5121/ijdms.2010.2401

and to solve domain problems [10]. Expert system is a computer tool that consists of user, knowledge base, and rules-based interface to response to user's query [11]. The main challenge in development of expert system described by Okafor and Osuagwu [12] is being the "processes of eliciting and representing knowledge". Knowledge Representation (KR) is a critical aspect of expert systems because it affects the implementation and performance of the system. To build an efficient expert system, a knowledge based structure must be the foundation that represents the domain's various knowledge types [13]. Furthermore, stages of knowledge use (Acquisition, Retrieval, and Reasoning) must be identified [14].
With more data being captured, need for information and analysis has exponentially grown, now users at all levels of an organization use reports and seek data to make decisions or take actions. Report requirements usually lead to several groups of related reports or parameterized reports, portals or dashboards. In situations where most of the requirements are unique and specific will lead to several discrete individual customs reports that may require increased report development, distribution and management time. As the volume of ad hoc reports increase valuable business information and knowledge is lost. We looked at reports with complex business rules, which queried multiple data sets that required complex query logic and developed an expert query system which introduces elements of standardization to shorten development time, to harness knowledge and improve manageability and scalability.

Section 2 outlines the importance and benefits of expert query systems and data queries. Section 3 presents the architectural layers of the expert query system to support the zone of standardization. Section 4 demonstrates the system development to harness knowledge from ad hoc queries to create the zone of standardization. Conclusions drawn based on this research project are discussed in Section 5.

2. EXPERT QUERY SYSTEM

Traditionally, ad hoc queries are issued directly against the data source or a copy with the same structure residing on a different system. Alternatively, using multidimensional database engines, data can be restructured and imported into specialized DBMSs to be optimized for queries [10]. Another approach is focused on limited data transformation such as partitioning data across multiple systems to access copy of data to take advantage of hardware parallelism [11]. In 1998, Nigrin and Kohane described source data query by defining "atomic queries" to simplify complex data retrieval for the non-programming clinicians [7]. A visual query model was developed by [13, 14] to support the discovery and distribution of data.

Generally, each ad hoc report maintains its own development path because of the specific nature of requirements, business rules, query logic, output and documentation. The discrete nature and redundancy of processing ad hoc reports may lead to inefficiencies in time, resource and valuable information. As the volume of ad hoc reports grow valuable business information and knowledge is being lost. However if the ad hoc queries are viewed within a context of larger collection of queries, we may be able to derive meaningful information worthy of reuse. Expert Query System (EQS) is an approach (Figure 1) that drives ad hoc queries towards standardization, structure and documentation to reuse, recycle queries, harness knowledge and reduce the foot prints of the subsequent ad hoc queries.

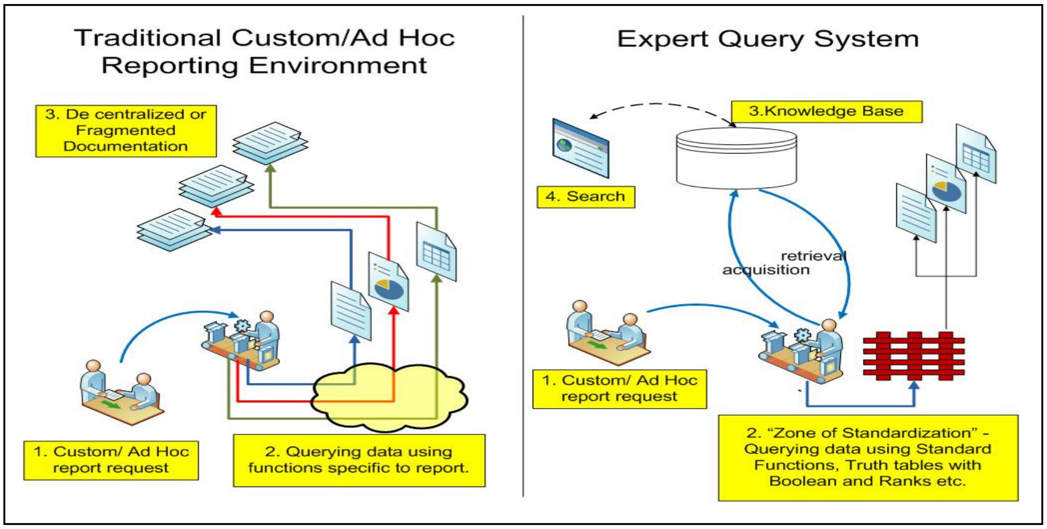

Traditional Custom/Ad Hoc Reporting Environment
3. De centralized or Fragmented Documentation
1. Custom/ Ad Hoc report request
2. Querying data using functions specific to report.
Expert Query System
3.Knowledge Base
4. Search
retrieval
acquisition
1. Custom/ Ad Hoc report request
2. "Zone of Standardization" - Querying data using Standard Functions, Truth tables with Boolean and Ranks etc.

Figure1. Traditional Custom/Ad Hoc Reporting Environment vs. Query Expert System

At the core of EQS is a "Zone of Standardization" (ZoS) which has two aspects, firstly it includes standardized code structure and secondly a standardized output in the form of truth tables with an indicator for each element of the result set indicating whether a particular element is a member of the set/subset or not. The extent of ZoS depends on common ground among ad hoc queries in the collection. The more similarities among the queries, the more expansive are the ZoS. ZoS also serves as a platform allowing knowledge acquisition of business and technical information. A search function on the knowledge base could return meaningful information and allow for reasoning. This mainly prevents re-inventing the wheel for the new ad hoc reports saving time and effort.

3. EQS ARCHITECTURE

We propose architecture of EQS as illustrated in Figure 2 based on four layers 1) Interface Layer exposes the interface of query system and allows each user request to automatically become accessible to end users. It provides the business users and technical users the ability to create, retrieve and update business and technical definitions. Manageability is a key function that is necessary for constantly changing business requirements; this layer is the gateway to the EQS, 2) Business Layer captures business definitions, technical definitions and relationships and creates the knowledge base. In the mean time, knowledge retrieval is done through the Interface layer. This layer delivers a high value of domain intelligence by enabling relevant information to decision makers as quickly as possible. It supports reporting and analytical capabilities by consolidating different source of data across the business domain, 3) Application Layer holds query library which contains the query code for the business definition and development platform for developing ad hoc queries. Most of the standardization and query development occurs in this layer. Knowledge base Application also sits in this layer and points to business definition, query meta-data and query library. This layer is the core of the system. It consists of the database applications, knowledge base applications, and user's APIs. It provides various query options, instance pooling of knowledge bases, and management of the distribution of tasks across computers, 4) Data Layer spans over data sets, database management systems and domain knowledge base from which the data is queried for the business requirements. This layer is responsible for storing data in a dedicated database. It includes a domain knowledge base and databases such as SQL Servers, flat files, Oracle, MySQL, DB2 and so on.

Business Layer and Application Layer together create the Zone of Standardization by driving requirements, definitions and query logic towards structured input and output via standardized functions and truth tables.

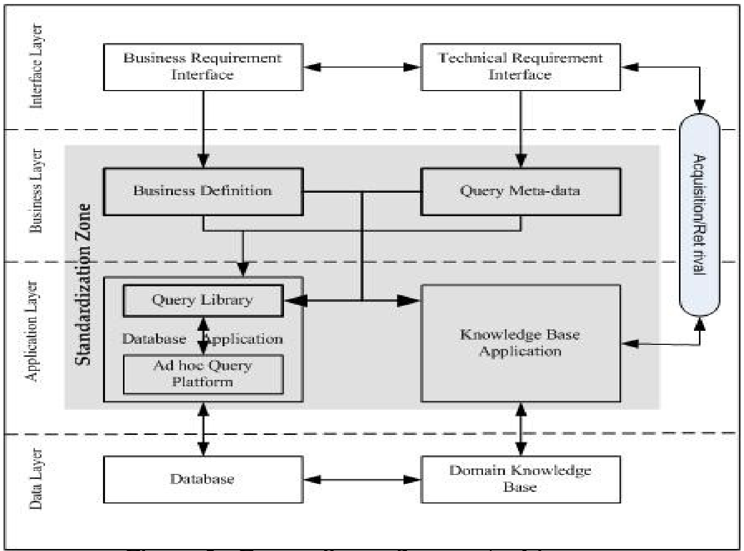


Figure 2. Expert Query System Architecture

The knowledge base application is responsible for knowledge acquisition and retrieval at a certain level of abstraction. Knowledge base applications use domain knowledge and databases to process basic data items by combining and sharing information captured by several different queries' requests.

The efficiency and performance of an expert system relies on the knowledge representation and knowledge use. The knowledge representation structure must include objects, instances, relations, and meta-knowledge. Furthermore, knowledge use stages (Acquisition and Retrieval) must be considered as presented in Figure 1 [9]. Holsapple and Joshi (2003) suggest that there are two level of acquisition of knowledge: one is the identification of existing knowledge and the other; the selection of desirable knowledge. These two activities requires effort and costs to structure facts in databases.

The retrieval technologies enable the database management system to store and manage explicit knowledge. However, retrieval of tacit knowledge usually requires collaborative computing systems such as intelligent agents, artificial intelligence, Extensible Markup Language (XML), and knowledge discovery in databases. The EQS Development Life Cycle (SDLC) is a knowledge intensive process that requires dynamic access to data, information, knowledge to support technical and business requirements. It ensures that all functional and user requirements and strategic goals and objectives are met. Furthermore, it provides a structured and standardized process for all phases of the EQS development effort to track the development of a system through several development stages from feasibility and requirements analysis, planning, development, testing, and deployment as presented in Figure 3. The agility of this life cycle is achieved based on knowledge that is acquired from real life data operation to inform functional and user requirements.

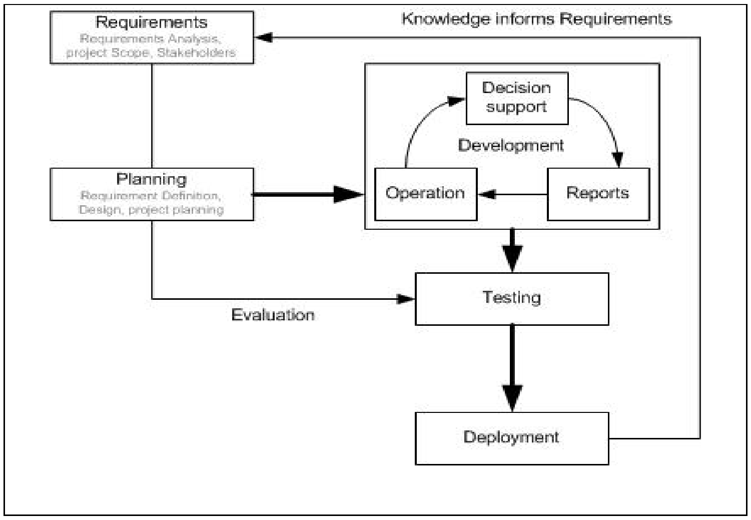


Figure 3. Expert Query System Development Life Cycle

4. THE ZONE OF STANDARDIZATION

The zone of standardization is achieved by passing standardized parameters into pre-built functions which then return standard values (1 or 0) to the result set table. The table values reflect the summary and breakdown, by Boolean indicator, whether a particular element is a member of the set/subset or not. This structure allows the business to perform the following operations on the data: set, bitwise, logical and statistical functions.

In other words, each row of the raw data set is examined to determine whether it meets the criteria or not. If it meets the specified criteria, then the function returns a standard value of one (1); if not, it returns a standard value of zero (0). In other words, the functions select (from the database) the count of records found, for this key, for these criteria. If the count is greater than or equal to one, the functions return one; else if not found, the functions return zero.

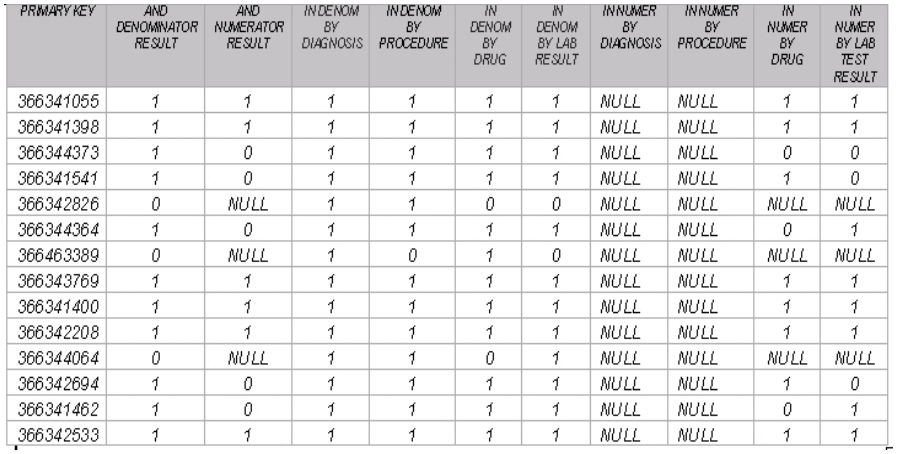
For business requirements, this concept provides the capabilities to define categories and subcategories for decision support investigation, define and associate variances for the categories/subcategories, define interval endpoints (such as dates and measurements), define and populate a code library having components (which define set inclusion/exclusion criteria) define and create standardized data operations, associate the categories/subcategories with the code library component(s) and also with the standardized data operations, append custom query criteria; all for singular or shared use.

For operations , this concept provides a business with the capabilities to dynamically interrogate the metadata defined in the requirements phase (above), detect the data operations and their pre-defined, pre-associated code segments, invoke the code associated with the data operations, and to populate a result set table.

For decision support, the result set table contains the summary and breakdown, by Boolean indicator, whether a particular element is a member of the set/subset or not. This structure allows the business to perform the following operations on the data: set, bitwise operations, logical operations and statistical.

For reporting, results may be grouped and rolled up by data element category, for example; by department, by physician, by patient, by medical intervention category, by medical result category, by date window

Table 1. Decision Support Data.



PRIMARY KEY	AND DENOMINATOR RESULT	AND NUMERATOR RESULT	IN DENOM BY DIAGNOSIS	IN DENOM BY PROCEDURE	IN DENOM BY DRUG	IN DENOM BY LAB RESULT	IN NUMER BY DIAGNOSIS	IN NUMER BY PROCEDURE	IN NUMER BY DRUG	IN NUMER BY LAB TEST RESULT
366341055	1	1	1	1	1	1	NULL	NULL	1	1
366341398	1	1	1	1	1	1	NULL	NULL	1	1
366344373	1	0	1	1	1	1	NULL	NULL	0	0
366341541	1	0	1	1	1	1	NULL	NULL	1	0
366342826	0	NULL	1	1	0	0	NULL	NULL	NULL	NULL
366344364	1	0	1	1	1	1	NULL	NULL	0	1
366463389	0	NULL	1	0	1	0	NULL	NULL	NULL	NULL
366343769	1	1	1	1	1	1	NULL	NULL	1	1
366341400	1	1	1	1	1	1	NULL	NULL	1	1
366342208	1	1	1	1	1	1	NULL	NULL	1	1
366344064	0	NULL	1	1	0	1	NULL	NULL	NULL	NULL
366342694	1	0	1	1	1	1	NULL	NULL	1	0
366341462	1	0	1	1	1	1	NULL	NULL	0	1
366342533	1	1	1	1	1	1	NULL	NULL	1	1

The decision support phase, the result set table contains the summary and breakdown, by Boolean indicator, whether a particular element is a member of the set/subset or not.

The result set table takes a form similar to a bit array. Standard columns exist regardless of whether they are utilized or not. In this example, diagnoses and procedures are not relevant for the numerator set. Therefore these columns are ignored. Each cell value is either 1 or 0 or null depending upon whether the criteria are met for that row, or null if not evaluated. In the example displayed below, after adding the not null denominator criteria; if the denominator result is 1, then the numerator criteria is evaluated. If the denominator result is 0 (not in the set) the evaluation of the numerator criteria is not executed.

In this study we looked at reporting landscape of a major health care organization in U.S. with about 900 beds, and 1.5 million clinic visits a year. Informational needs of this organization were great an average of 120 ad hoc reports were being produced per month. We looked at set of discrete queries, each with their own criteria. We standardized the logic using functions as shown below. Each query logic is wrapped in a standard function that returns excluded 'E' or '0' from the set or included 'I' or '1' based on the business criteria.

```
Select
Count(*)
into v_count
from
source tables

/* query logic goes here..*/
….
If v_count>0
then return '0'; -- exclusions
else
return '1' -- inclusions
end if;
end;
```

In table 1, we are checking numerator and denominator for four criteria diagnosis, procedure, drug, and lab. Numerator is a subset of denominator. The first column in the table is the identifier column, the next two columns "AND DENOMINATOR RESULT" and "AND NUMERATOR RESULT" are the summary columns based on which decisions are made, and they are computed based on the results of all denominator and numerator columns. The following four Colums prefixed with "IN DENOM" check for the denominator criteria, the last four columns prefixed with "INNUMER" check for the numerator criteria.
For rows having keys *366342826, 366463389* and *366344064* those rows are not in the denominator set, therefore the numerator criteria are not evaluated as illustrated in Table 1.
The illustrated example is widely used to make decision in health care to measure quality of care, performance etc. The emerging research area of health and medicine has major initiatives to improve data and information quality and accuracy. The Agency for Healthcare Re-search and Quality (AHRQ) as part of the US Department of Health Services (HHS) is promoting use of health information technology to improve the quality health care. [1].

This solution was developed using relational database and procedural language for a Health System decision support system. It could also be used for any business application having the requirements to define sets and subsets, and to populate the sets according to defined criteria. Once the final result sets are populated, mathematical and statistical functions, set operations, and reporting may be performed in a standardized manner.
Tables contain:

1. Definitions of n-level (nested) business categories. The definition of nested categories allows another level of information to be presented in the report such as multiple levels from one dimension, levels from different dimensions. For example an organization can apply nested category to sales regions under product.
2. Standardized metadata allows users process more information from sources and formats to improve information reusability and supports internal processes of an organization (example date intervals, codes, criteria).
3. Offset variances for endpoints of standardized metadata.
4. A library of code segments for singular or shared use.
5. Standardized data operations and consistant data understanding enhances data interoprtability. The methods that make up metadata are known as constraints or stereotyped operations. They depend on the requirement to map the external service methods or database queries to standardized data operations. For example, The EQS supports stereotyped operations such as create, retrieve, update, or delete an item.
6. Relationships between the metadata, the code library component(s) and the standardized data operations. Standardized operations include:

1. Dynamic interrogation of the metadata.
2. Dynamic detection of the data operations and their pre-defined, pre-associated code segments.
3. Dynamic invocatin of the code associated with the data operations.
4. Dynamic population of a result set table.

The result set table contains the summary and breakdown, by Boolean indicator, whether a particular element is a member of a set/subset or not. This structure allows the business to:

1. Perform set operations on the data to combine set of rows returned by a query such as union, intersect, and minus.
2. Perform bit operations on the data.
3. Perform logical operations on the data.
4. Perform statistical operations on the data.
5. Create reports and charts via a standardized method (all result set tables are of the same format).
6. Group and rollup by data element category. For example, by department, by physician,

by patient, by medical intervention category, by medical result category, by date window.

5. CONCLUSIONS

We have presented a knowledge-based EQS framework that drives ad hoc queries towards standardization, structure and documentation, to re-use, recycle queries and harness knowledge. This framework is focused on capturing elements of business definitions and utilizing Boolean indicators for each element of the result set.

This approach resulted in a reduction in lines of custom code required to create one report (or one result set table in our design) from approximately 8000 lines of single-use code to approximately 2300 lines of shared-use code in the code library. It is a knowledge-based system that increases visibility into key metrics by allowing domain users to meet their reporting needs with complete self-service reporting and ad hoc query capabilities. It provides managers and business users the capability to access, modify, or author reports quickly and easily. Expert Query System (EQS) is an approach that drives ad hoc queries towards standardization, structure and documentation to reuse, recycle queries and harness knowledge to reduce the foot prints of the subsequent ad hoc queries.

ACKNOWLEDGEMENTS

The authors wish to express their appreciation to the sponsors of the EQS project for providing an opportunity to test a large set of data and for their encouragement with this research.

REFERENCES

[1] Agency for Healthcare Research and Quality (AHRQ), http://www.ahrq.gov/RESEARCH/hitfact.htm, Accessed 20 January.

[2] Jackson, P. “MYCIN: Medical Diagnosis using Production Rules”. Introduction To Expert Systems. Addison-Wesley Publishing Company, (1986):(102-114).

[3] Turban, E., & Aronson, J. E. (2001). Decision support systems and intelligent systems, sixth Edition (6th ed). Hong Kong: Prentice International Hall.

[4] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. Communications of ACM, 39, 27–34.

[5] Veloso, M. & Borrajo, D (1994). Learning strategy knowledge incrementally. Tools with Artificial Intelligence. Proceedings of Sixth International Conference (pp. 484–490).

[6] Giarratano, J.C., Riley, G.D. “Introduction To Expert Systems”. Expert Systems 4thEdition. Thomson Course Technology, (2005):(19,29,32,34).

[7] Nigrin D, Kohane I. Data mining by clinicians. Proc AMIA Annu Fall Symp. 1998:957–61.

[8] Fiegenbaum, E. A., Barr A., and Cohen, P. R. (eds) (1981-1982). Handbook of Artificial Intelligence, Vol. 1-3, HeurisTech Press, William Kaufmann Inc., Stanford, CA,.

[9] Firebaugh Morris W. (1998). *Artificial Intelligence: A Knowledge-Based Approach*, PWS-KENT Publishing Company, Boston, pp. 275-303.

[10] Thomsen E. OLAP Solutions: Building Multidimensional Information Systems. New York: Wiley, 2000.

[11] Inmon W, Rudin K, Buss C, Sousa R. Data Warehouse Performance. New York: Wiley, 1998.

[12] Eric C. Okafor, C.E., Osuagwu, C,C. (2007). Issues in Structuring the Knowledge base of Expert Systems. Electronic Journal of Knowledge Management (EJKM), Volume 5 Issue 3, pp 313 – 322. http:// www.ejkm.com.

13] Murphy SN, Barnett GO, Chueh HC: Visual query tool for finding patent cohorts from a clinical data warehouse of the Partners HealthCare System. *J Am Med Inform Assoc* 2000:1174.

[14] Murphy SN, Chueh HC: Visual query tool for integrating clinical and genetic data in the in the Partners Healthcare System. *J Am Med Inform Assoc* 2001:983.

[15] Holsapple, C. W. and K. D. Joshi, "A Knowledge Management Ontology," in *Handbook on Knowledge Management*, Volume 1k (ed.C. W. Holsapple). New York: Springer-Verlag, 2003, pp. 89–128. InfoWorld, "Enterprise Knowledge Portals Wise Up Your Business,"*InfoWorld,* 22(49), December 4, 2000, *archive.infoworld.com/gui/infographics/00/12/04/001204tcvit.gif* (accessedJuly 2003).

Authors

Jehad S. Alomari: *assistance professor of CIS/MIS* with Taif University, Taif, KSA, 21974.Doctorate of management in Information Technology. Lawrence Technological University (2009). MS (CIS) Eastern Michigan University (2003). I have a strong background in diverse fields such as Software Engineering, Enterprise Architecture, Software Design, Knowledge Engineering, and ontology development.

Farrukh T. Mohammed; MBA (2007), MS (GIS) (2006); University of Michigan Health System (2001) and has a strong background in Decision Support Systems, Business Intelligence, Predictive Analytics and Enterprise Reporting and Analytics in Health Systems.

Kathleen E. Hayden; BS (Computer Science); University of Michigan Health System (2001) and has a strong background in Relational Database Systems, Database Administration and Performance improvement.