Integrating a Modern Knowledge-Based System Architecture with a Legacy VA Database: The ATHENA and EON Projects at Stanford

Aneel Advani, MD,[1] Samson Tu, MS,[1] Martin O'Connor, MSc,[1]
Robert Coleman, MS,[2] Mary K. Goldstein, MD,[2,3] and Mark Musen, MD, PhD[1]
[1]Stanford Medical Informatics and [2]VA Palo Alto Health Care System
[3]Stanford University School of Medicine
Stanford University, Stanford, CA 94305-5479

We present a methodology and database mediator tool for integrating modern knowledge-based systems, such as the Stanford EON architecture for automated guideline-based decision-support, with legacy databases, such as the Veterans Health Information Systems & Technology Architecture (VISTA) systems, which are used nation-wide Specifically, we discuss designs for database integration in ATHENA, a system for hypertension care based on EON, at the VA Palo Alto Health Care System (VAPAHCS). We describe a new database mediator that affords the EON system both physical and logical data independence from the legacy VA database. We found that to achieve these design goals, the mediator requires two separate mapping levels and must itself involve a knowledge-based component.

INTRODUCTION

We describe the function of a database mediator that allows a modern component-based medical decision-support system (DSS), the Stanford EON architecture,[1] to use a legacy database, the VA VISTA database.[2] The VA VISTA system (which is also referred to by its older acronym DHCP) occupies a unique position as a nation-wide electronic clinical medical record. It offers excellent potential for the deployment of advanced DSSs in order to effect changes to patient care in real-world clinical settings. However, one of the major hurdles for clinical application is that a DSS be able to use the legacy database infrastructure available in the VA clinical environment.

The problem is that frequently there are conflicts between the data models of legacy database systems and the data models assumed by the problem solving methods in the DSS.[3] Moreover, the legacy databases such as the VISTA in the VA system often have heterogeneous and inconsistent schemas that have accumulated over time. Lastly, some legacy systems, such as the VA VISTA system use a hierarchical database structure, which presents unique challenges for those DSSs that assume relational databases. These challenges include the lack of a high-level query language standard and the more limited expressiveness of the hierarchical database model.

The ATHENA Application

One of the clinical applications that uses the EON decision-support architecture at Stanford is the ATHENA (Assessment and Treatment of Hypertension: Evidence-Based Automation) project at the VAPAHCS. The ATHENA program is an evaluation of clinical practice guidelines implemented using an automated DSS. The overall task in the intervention is to improve compliance with national hypertension care guidelines.[4] This DSS is to be deployed in 11 different clinical sites for greater than 7000 patients who are part of the primary care population of the VAPAHCS. The ATHENA system is one of the first automated knowledge-based DSS applications for guideline-based medical care in the VA system.

The ATHENA project uses the component-based EON architecture developed at Stanford to automate certain tasks associated with guideline-based care. For each task, there is a specific reasoner or *problem-solving module* (PSM) corresponding to a component

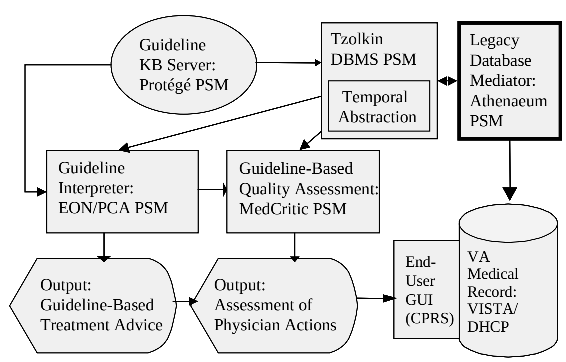

Figure 1 ATHENA system incorporating the EON architecture for component-based decision-support. Each EON component carries out a specific problem-solving task for automated decision-support of guideline-based care.

in the EON architecture to carry it out. Such tasks include the monitoring the execution of an applied guideline, using the EON Protocol Compliance Advisor (PCA) PSM and assessing the quality of the guideline-based treatment, using the MedCritic PSM (see Figure 1).

THE ATHENAEUM MEDIATOR

All the components of the EON architecture use a temporal database management system, called the Tzolkin module, to process queries to an EON-compatible relational database.[5] However, in the case of an actual hospital environment, this standard EON relational data model must be reconciled with the data model used in the VA medical record database. We have designed a knowledge-based mediator, called Athenaeum,[†] to accomplish this integration. The Athenaeum mediator maps one database to another based on the separation of data models into three layers by the ANSI Standards Planning and Requirements Committee (see Table 1).[6] This methodology allows us to divide the integration of the legacy database into two separate mapping steps.

First, the conceptual schema of the legacy database is mapped to an intermediate relational schema. This step insulates the application's view of the legacy from the conflicting physical design of the legacy database, a concept known in the database literature as *physical data independence*.[3] Second, the intermediate relational schema is mapped to the application-specific external schema, in this case the EON data model. This insulates the application's view of the legacy database from the mismatched semantics of the conceptual schema in the legacy database. This concept is referred to as *logical data independence*.[3] Note that we don't actually create a redundant intermediate relational database, only an intermediate relational schema. The separate mapping allows us to isolate the physical independence from the logical independence of the mapped legacy database.

To implement the integration, the Athenaeum mediator uses two different knowledge-bases (see Figure 2). First, there is a declarative "schema knowledge-base" that describes an internal semantic transformation from a physical legacy database (such as the hierarchical M-based VA VISTA system) to a relational mapping of its contents. Second, the Athenaeum module contains a "mapping knowledge-base" that maps the relational view of the legacy database to the external database-independent relational data model assumed by the EON problem solvers. The Athenaeum mediator uses the schema and mapping knowledge to provide the correct results of selection and projection operations during queries by the EON module to the data management module, Tzolkin.

We will use the following example EON Tzolkin query to illustrate the use of the schema and the mapping knowledge-bases in the semantic transformation from the VA data in order to satisfy the data requirements of the EON PSMs. The query asks when was the last time interval during which patient with id=012345 had primary hypertension that was well controlled for more than six months and during which the patient has a creatinine value of less than 2.1. The query is written in Chronus II, an extended SQL language used in EON to perform temporal abstractions.[7] It has the form:

```
TEMPORAL SELECT
        LAST problem-name, problem-value
FROM patient-problems-view AS TI
        patient-labs-view AS T2
WHERE problem-name = "HTN"
    AND problem-value = "Well_Controlled"
    AND lab-name = "Cr"
    AND lab-value < 2.1
    AND patient-id = "012345"
WHEN DURATION(T1) > MONTHS(6)
    AND CONTAINS(T1,T2)
```

Note that the query presupposes the EON relational model on which to carry out its temporal select, projection, and temporal join operators.[7] Since the VA database has a hierarchical internal schema, the Athenaeum mediator first maps this to an intermediate relational conceptual schema.

Table 1 Classification of 3 levels of data models into internal, conceptual, and external schema. Athenaeum maps the VA FileMan schema into an intermediate relational schema, and then maps this relational schema to the EON Data Model.

ANSI/SPARC 3-level data model	Database design terminology	Legacy VA database	EON relational database
1. Internal Schema	Physical Design	Hierarchical Files	Relational Tables
2.Conceptual Schema or Logical Data Model	Specific to Database System Implementation	VA FileMan Database System	EON Relational Schema
3. External Schema or View and Query Level	Independent of Database but Specific to Application	M Modules (Procedural - not a Data Model)	EON Data Model and Chronus II Language

[†] After the name for the temple of Athena, or "house of wisdom".

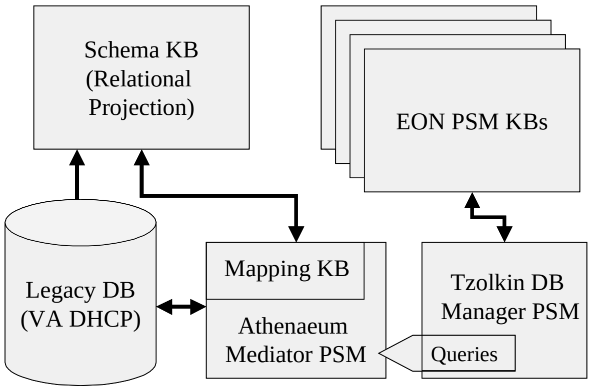

Figure 2 The Athenaeum mediator consists of three parts: (1) a schema knowledge-base, (2) a mapping knowledge-base, and (3) a database mediator PSM.

HIERARCHICAL TO RELATIONAL MAPPING USING SCHEMA KB

The VA VISTA system was developed in the 1970s based on a centralized hierarchical database, called FileMan. FileMan is accessed through the M (MUMPS) language.[8] The hierarchical model, rather than the current relational model represented a standard at the time the VISTA system was first developed. The main characteristic that distinguishes the hierarchical model from the more modern relational model is its restriction to supporting only one-to-many relations between record-sets, or *files* in the VA VISTA notation.[6] This model stands in contrast to the relational approach that can support many-to-one, and many-to-many relations between tables directly. Hierarchical databases cannot support these types of relations without computationally expensive work-arounds involving record duplication and redundancy.

For example, to find all the creatinine measurements over multiple patients may be much harder in the hierarchical database. Since the Lab_Data file is a child of the Outpatient Encounter file, which depends on a particular patient encounter, a FileMan search on the Lab_Data file for “all the patients with creatinine < 2.1” would be require procedural M coding. The relational model would do this with a simple SQL query on the Current_Labs_and_Vitals table. Thus, the increased expressiveness and compactness of the relational model strongly biases creators of knowledge-based systems to use it as the application data model. Hence the need for a hierarchical to relational mapping mechanism.

The Athenaeum module uses the schema knowledge-base to accomplish this mapping from the physical design of the hierarchical database to a relational schema. This knowledge-base, built using the Protégé knowledge server tools,[1] is used to map the relevant hierarchical data files in FileMan to a set of relational tables (see o relational mapping mechanism.

The Athenaeum module uses the schema knowledge-base to accomplish this mapping from the physical design of the hierarchical database to a relational schema. This knowledge-base, built using the). For our example query, the information on the creatinine value would have to be obtained from the Lab_Data file and mapped to the Current Labs_and_Vitals table. The following frame, entered in our schema knowledge-base, allows this mapping:

Target: Current_Labs_and_Vitals.Creatinine
Target-type: DECIMAL(4,2)
⇐ CHAR_TO_FLOAT, ALL_ENTRIES
Source: File 63, SubFile 04, SubFile 4
Source-type: VARCHAR(6)

The target specifies that the Creatinine field for Current_Labs_and_Vitals holds data in the DECIMAL format from the VA File 63.04.4 which has the creatinine values in the form of characters. The transformations needed include the type change from character to float, and the choice to import all values of creatinine currently in the VA file. Note that this mapping also allows us to import controlled terminology of drugs and diagnoses from the associated files of the VA VISTA database.

Even in this first mapping stage, however, it is clear that several semantic choices have to be made with respect to the relational mapping. For instance, our sample query requires identifying primary hypertension in the patient. However, there were two

Table 2 Relational mapping of the hierarchical VA data model. Column on the left indicates the hierarchical files used to create the relational tables, which are listed in the column on the right.

Hierarchical Model VA FileMan Files	Tables in the Relational mapping of VA Files
• Hospital_Location • Outpatient_Encounter	Current_Encounter – date, location, provider
• Purpose_of_Visit • Outpatient_Diagnosis	Active_Problem_List –noted at time of visit
Outpatient_Prescription	Current_Prescriptions – both active and allowed
• Lab_Data • Vitals_Measurement	Current_Labs_and_Vitals – inpatient and outpat.
• Patient File	Pt_Demographics
• Drug_File	Drug Names Vocab.
• ICD_Diagnosis	Diagnosis Codes/Names
• Labs Data Dict	Lab Tests Field Codes

different potential sources for the current outpatient diagnoses for a given patient: the Purpose_of_Visit file and the Problem_List file. The Purpose_of_Visit file was used for identifying the chief complaint and for billing the visit, while the Problem_List file contained a longitudinal problem list that must be kept up to date by primary care physicians. A chart review of 148 charts showed that the Purpose_of_Visit file was 100% sensitive for hypertension, and 79% specific when the chart was used as a gold standard. The Problem_List was more specific at 95%, but less sensitive at 65%.[9] We made the choice to accept some false positives in order to identify more true positive HTN patients. Thus, in the ATHENA project, a specific choice was made to point to the Purpose_of_Visit file for identifying patients with hypertension in the schema knowledge base. These semantic mismatches in the "raw" physical files made it necessary to use a configurable knowledge-based approach in creating the relational mapping of the hierarchical database. Hard-coding this mapping would not ensure physical data independence as the national schema recommendations evolved over time.

EON DATA MODEL MAPPING

In the previous mapping, we were successful in removing the physical data dependence in the semantics of the VA legacy database. However, there are additional semantic mismatches which exist at the conceptual or logical level of the ANSI hierarchy presented in Table 1.[3] These include mismatches in scope restriction, domain semantics, temporal basis, temporal granularity, and temporal abstraction. Therefore, the Athenaeum mediator includes a mapping knowledge base to create another set of relational tables. These tables, in addition to being internally consistent, are also semantically consistent with the EON data model. Table 3 shows how each of these semantic mismatches occurs in the example query presented above.

For instance, the query looks for records where 'lab-name "Cr" AND lab-value < 2.1'. The VA Lab_Data file, and hence also the VA Current_Labs_and_Vitals table, include both inpatient and outpatient laboratory results. However, the EON outpatient Protocol Advisor does not filter out values that are not outpatient measurements. Knowledge to accomplish this filtering is not embedded anywhere in the database itself. This *scope restriction* mismatch is resolved by encoding the following frame in the mapping knowledge-base:

Target: Patient_Labs_View.Creatinine
$\Leftarrow$ SCOPE_RESTRICTION(Filter_Outpatient)
Source: Current_Labs_and_Vitals.Creatinine

where Filter_Outpatient is a scope restriction function. It is evaluated by the Athenaeum module using a Chronus II query to the Tzolkin module:

```
SELECT Current_Labs_and_Vitals.Creatinine
FROM Current_Labs_and_Vitals AS T1,
        Outpatient_Encounter AS T2
WHEN CONTAINS (T2,T1)
```

This scope restriction must be used so that we do not use inpatient laboratory data for reasoning about guidelines that only consider outpatient diagnoses. The interpretation of a creatinine value > 2.1 might change in the context of an acute hospital illness.

Similarly, the first condition in the WHERE clause, 'problem-name = "HTN"' refers to a problem called "HTN" and not to an ICD-9 code. The diagnostic information from the Purpose_of_Visit file in the VA database comes in the form of ICD-9 codes. Thus, the use of "HTN" as a well-defined problem name depends on a knowledge base that transforms the domain semantics of the "raw" diagnostic codes into well-formed elements of a patient problem list. In this case, our knowledge base used the ICD-9 codes 401.1 and 401.9 for primary and unspecified hypertension respectively as comprising the patient problem of HTN. This excluded the codes 401.0 for malignant HTN and 402.x for HTN with concomitant heart disease from the category. The particular domain semantic transform chosen for the mapping knowledge-base was dictated by the inclusion criteria set out by the national standard guideline for the management of primary HTN.

The reasoning steps described above require a separate application-specific mapping knowledge-base and a mediator that can use the mapping knowledge. With the separation of the schema

EON Module	EON Data Query	VA Data Model	Semantic Transform
Protocol Advisor	Cr from Outpt Labs	Cr from Inpt or Outpt file	Scope Restriction
Protocol Advisor PCA	HTN Problem Name	ICD-9 Code 401.1	Domain Semantic Transform
MedCritic	Transaction Time of Cr	Valid Time of Cr	Temporal Basis
Tzolkin DBMS	Time Grain of Month	Timestamp Day/Time	Temporal Granularity
Tzolkin DBMS	HTN = Well Controlled	BP values over time	Temporal Abstraction

Table 3 Examples of semantic transformations required from VA data model to EON data model that are required to answer the example query.

semantic transformations in the application-specific mapping to ensure logical data independence knowledge-base mapping we can use the Chronus II query language on the intermediate relational schema. This allows us to enforce more complex.

DISCUSSION

The functions supported by the Athenaeum mediator include schema integration, semantic conflict resolution, and vocabulary/data dictionary (meta-data) integration. However, unlike some other database mediator architectures,[10] that use a generic relational model of medical care, our approach uses an application-specific knowledge-base optimized to each guideline task performed by components in the EON architecture. We also use the Chronus II query language to find the results of semantic transformations that must be applied as a result of the mapping knowledge.

This approach allows us to avoid the pitfalls of overly general approaches to database mediation. A case in point is the standard relational mapping of the VISTA system, the SQL Interface (SQL-I) system. [11] As we saw in the discussion above, the more than 400 record-sets or files that make up the VISTA system frequently contain inconsistencies in semantic scope and conflicting assumptions in domain semantics. To use the SQL-I interface as the only relational model for the VA databases would be quite inadequate. Moreover, to actually operate on the VA databases, one has to use an additional proprietary M-to-SQL product. These mapping have their own proprietary semantic assumptions about the most appropriate mapping from the VA database to a relational model. Our knowledge-based approach allows the application-specific knowledge to be separated from the question of relational mapping. Moreover, the schema knowledge-base also removes the internal semantic inconsistencies which the SQL-I relational mapping simply propagates. This allows us to achieve true physical data independence that cannot be achieved using the combination of SQL-I and a proprietary M-to-SQL product.

Thus, a straight generalized relational mapping, without a properly constructed semantically consistent knowledge base cannot be used for automated decision-support. Our approach allows a knowledge-based mediator to represent and automatically resolve inconsistencies and semantic domain constraints that can only be detected by human experts familiar with the guideline application and the assumptions of the VISTA system. Secondly, since these knowledge-bases are configurable, when the underlying VA database change in some way, only the knowledge-bases and not the mediator code itself need to be updated.

Acknowledgements

This work was supported in part by NLM grants LM05708 and LM06245 and VA grant CPG-97-006. We thank Larry Lou for programming assistance on this project.

References

[1] Musen MA, Tu SW, Das AK, Shahar Y. EON: A Component-Based Approach to Automation of Protocol-Directed Therapy, Journal of the American Medical Information Association1996;3(6):367–388.

[2] Beattie, MC, Buxton, EC, Wiederhold, V. VA Databases Resource Guide, Vol. V, DHCP V3.0. Palo Alto: VAPAHCS, HSR&D Center for Health Care Evaluation; 1996.

[3] Elmagarmid A, Rusinkiewicz M, Sheth A, eds. Management of Heterogeneous and Autonomous Database Systems. San Francisco: Morgan Kaufmann; 1999.

[4] National High Blood Pressure Education Program. The Sixth Report of the Joint National Committee on Detection, Evaluation, and Treatment of High Blood Pressure. Washington: NIH; 1998.

[5] Nguyen JH, Shahar Y, Tu SW, Das AK, Musen MA. A Temporal Database Mediator For Protocol-Based Decision Support. 1997 AMIA Annual Fall Symposium, Nashville, TN, 298-302. 1997.

[6] ANSI Study Group on Data Base Management Systems, Interim Report. Bulletin of ACM SIGMOD (FDT); 1975; 7(2).

[7] O'Connor MJ, Tu SW, Musen MA. Applying Temporal Joins to Clinical Databases. Stanford Medical Informatics, Technical Report, Stanford University School of Medicine, Stanford, CA; 1999.

[8] VISTA Software Development Team, Dept. of Veterans Affairs. VA FileMan v. 21.0 Programmer Manual; 1997. URL:www.vista.med.va.gov/softserv.

[9] Szeto H, Goldstein MK. Accuracy of Computer-Identified Diagnoses in a VA Medical Clinic. Presented at the 17th Annual VA HSR&D Service Meeting, Washington D.C. Feb 24-26, 1999

[10] Sujansky W, Altman RB. An Evaluation of the TransFER Model for Sharing Clinical Decision-Support Applications. In James J. Cimino, Ed., Proceedings of the 1996 AMIA Annual Fall Symposium, Washington, D.C.;1996; 468-472.

[11] VISTA Software Development, Dept. of Veterans Affairs. VA FileMan SQL Interface (SQLI) Site Manual; 1997. URL:www.vista.med.va.gov/softserv.