Model Formulation ■

Automated Database Mediation Using Ontological Metadata Mappings

Luis Marenco, MD, Rixin Wang, PhD, Prakash Nadkarni, MD

Abstract **Objective:** To devise an automated approach for integrating federated database information using database ontologies constructed from their extended metadata.

Background: One challenge of database federation is that the granularity of representation of equivalent data varies across systems. Dealing effectively with this problem is analogous to dealing with precoordinated vs. postcoordinated concepts in biomedical ontologies.

Model Description: The authors describe an approach based on ontological metadata mapping rules defined with elements of a global vocabulary, which allows a query specified at one granularity level to fetch data, where possible, from databases within the federation that use different granularities. This is implemented in OntoMediator, a newly developed production component of our previously described Query Integrator System. OntoMediator's operation is illustrated with a query that accesses three geographically separate, interoperating databases. An example based on SNOMED also illustrates the applicability of high-level rules to support the enforcement of constraints that can prevent inappropriate curator or power-user actions.

Summary: A rule-based framework simplifies the design and maintenance of systems where categories of data must be mapped to each other, for the purpose of either cross-database query or for curation of the contents of compositional controlled vocabularies.

■ **J Am Med Inform Assoc.** 2009;16:723–737. DOI 10.1197/jamia.M3031.

Introduction

One challenge in federated database integration is that databases from various research groups may store information on the same category of data differently. *Physical* heterogeneity issues, such as different internal names for semantically equivalent tables/columns, data stored in a single table vs. multiple tables, data stored in column-modeled vs. row-modeled form, and so on, have been addressed successfully through standard approaches such as database views and mappings of individual database schema elements to a global schema. However, semantic representation differences—notably due to equivalent information being stored at different granularity—cannot be addressed using these mechanisms. For example, in one database, various attributes of a concept—e.g., morphology, location, tissue type—may be represented as distinct fields, while in another database these attributes may be combined implicitly through the descriptive name of that concept. This paper describes a general approach to specifying equivalence between different concepts when such granularity differences exist. We describe an implementation for integration of neuroscience databases, and provide another example in the biomedical controlled vocabulary domain. This work may lay the foundation for database-contextual information integration in biosciences and other areas of research.

Affiliations of the authors: Center for Medical Informatics (LM, RW), Department of Anesthesiology (LM), Yale University School of Medicine, New Haven, CT; Geisinger Health Systems (PN), Danville, PA.

This research is supported by NIH Grants R01 DA021253 and P01 DC04732.

The authors thank the curators of the CCDB and CoCoDat databases for making their data available for use in this paper, and Dr. Gordon Shepherd for curating the neuron ontology used in this work.

Correspondence: Luis Marenco, MD, Center for Medical Informatics, Yale University School of Medicine, PO Box 208009, New Haven, CT 06520-8009; e-mail: <luis.marenco@yale.edu>.

Received for review: 10/13/08; accepted for publication: 06/07/09.

Background

Approaches to Database Integration

Two general approaches to database integration are the warehouse approach, which emphasizes data translation, and the database federation approach, which emphasizes query translation.[1] In the warehouse approach, individual sources' data are converted to a common grain and moved to a read-only warehouse whose data model is the union of the source models. Query optimization, performance, and a priori data validation and curation are readily addressed, but this approach succeeds only with a high degree of central control, which is not seen in most collaborative research scenarios. It is also much less workable for research databases where schemas and metadata change constantly. In the federated database approach, only information about the individual sources' data models (schema metadata) is integrated centrally into a federated schema;[2,3] the data sources remain separate. "Mediator" software translates queries against the federated schema into queries against individual sources and merges the results. Differences in granularity, encoding, and representation semantics make

this approach challenging, in addition to limiting the quality and value of the output. While simple Web-browsing interfaces that show details of a single item of interest are feasible, representing sophisticated analytic queries involving complex Boolean logic that return numerous rows of data are rarely possible.

The Problem: Querying across Datasets of Heterogeneous Granularity

Successful query of a federated schema depends on being able to correctly identify—that is, "map"—equivalent concepts across different databases: varying data granularity across databases complicates the query formulation process. Fortunately, granularity variation is not entirely ad hoc: there exists a definite logical model underlying the representation of all of a given database's concepts. However, because this model is very often *implicit*, solving the mapping problem across databases requires creating an equivalent *explicit* representation.

Certain types of granularity heterogeneity are seen in databases that need to be integrated before metaanalysis.[4] Here, one encounters parameters of the *nominal* (enumerated) or *ordinal* (ranked) data type, whose valid values belong to a discrete set, but where the set members are defined differently for the same parameter in different databases. In this domain, it is well known that one can generally map a finer grain to a coarser grain but not vice versa. For example, if one dataset records "smoking status" as "Smoker/Non-Smoker", and another as "cigarettes/day", one can infer that anyone who smokes more than zero cigarettes per day is a smoker, but inferences in the reverse direction are not possible. Grain translations are achieved here through the straightforward approach of translation (lookup) tables that map the levels in one measure to levels in the other. While equivalent to if-then rules, table-driven methods have the advantage of being more efficient and easier to modify.[5]

The granularity problem described in this paper, which we believe has not been addressed previously in a systematic fashion, concerns *explicit* versus *implicit representation of concept attributes*. That is, attributes are represented explicitly as separate fields for a concept category in one database, but merged into the textual descriptions of those concepts in another database. We show that this problem is intimately linked to the issue of compositional versus noncompositional concept representation in controlled vocabularies, described below.

The Concept Representation Problem in the Biomedical Vocabulary Domain

To provide background for our approach for handling heterogeneity in concept attribute representation, we summarize a well-known issue in the controlled vocabulary field. There are two ways to combine new concepts from existing concepts as knowledge within a domain evolves. With *precoordination*, the vocabulary's *curators* create a new concept entry with a descriptive phrase that captures the meaning of a combination—e.g., "renal hypertension". With *postcoordination* (composition), the vocabulary's *power-users* combine existing concepts (here, "secondary hypertension" and "kidney disease") into a miniature semantic network using relationship edges (such as "Cause-of") from a set of permissible relationship types. The pros and cons of each approach are discussed by White et al.[6] With precoordination, while curatorial fiat weeds out nonsensical combinations, concepts can still proliferate. Tasks such as concept-based document searching and manual/electronic concept matching of text become much more complicated: highly complex concepts are very difficult to match exactly. A compositional vocabulary's contents are much less likely to proliferate, but a naive user may combine concepts in meaningless ways. To mitigate this risk, it is desirable to specify *computable* constraining rules for concept composition. However, to the best of our knowledge, no constraint-based computational framework is operational in the biomedical ontology domain.

- In the precoordinated vocabulary LOINC (logical observations, identifiers, names and codes),[7] a concept is defined based on a *fixed* combination of the parameter being recorded/measured, the property recorded, temporal aspects of the recording, the system in which the parameter was recorded (e.g., blood, urine), the recording/measuring method and the scale (data type) in which the result is expressed. That is, the compositional rule is *implicit* and *hard-coded* in the structure of LOINC's schema.
- With the compositional systematic nomenclature of medicine (SNOMED),[8] the composition of individual complex concepts in terms of others is recorded through pairwise relationships between a complex concept and one or more simpler concepts. Sometimes, a set of relationships is grouped together using an integer Relation Group field to indicate that they collectively form a single semantic unit. However, constraining rules that apply to entire *families* of concepts, such as the categories of data that a given attribute applies to, and the values allowed for that attribute, are defined *only in prose form* in the SNOMED user Guide.
- Finally, the web ontology language (OWL) is intended to support ontology development and interchange using an XML-based syntax. While OWL supports definition of certain constraints based on description logics,[9] the constraint syntax, which is fixed and nonextensible, does not support compositional rules to deal with concept-granularity heterogeneity.

Existing Approaches to Query Mediation

From the very extensive and diverse literature on query mediation, we summarize briefly previous approaches that either use metadata or ontologies to drive the query process, or those that focus on bioscience-related problems.

Several projects have explored database interoperation/integration issues in genomics.[10,11] BioMediator,[12,13] developed at the University of Washington, uses a federated approach to access a number of public genomics-related data sources, e.g., NCBI entrez,[14] Online Mendelian Inheritance in Man (OMIM),[15] and Gene Ontology.[16] The global data model treats both concepts and data as nodes in a semantic net. This work has subsequently been extended to the neuroscience domain.[17] In neuroscience, the UCSD BIRN (Biomedical Informatics Research Network) Database Mediator[18,19] uses a centralized data repository with remote views that collectively act as a global ontology. CaGrid[20,21]