2003

A PROCESS FOR ACHIEVING COMPARABLE DATA FROM HETEROGENEOUS DATABASES

Rachel L. Richesson
The University of Texas School of Health Information Sciences at Houston

A PROCESS FOR ACHIEVING COMPARABLE DATA FROM HETEROGENEOUS DATABASES

A

DISSERTATION

Presented to the Faculty of
The University of Texas
Health Science Center at Houston
School of Health Information Sciences
in Partial Fulfillment
of the Requirements

for the Degree of

Doctor of Philosophy

by

Rachel L. Richesson, MS, MPH, PhD Candidate[1]

Committee Members:

James P. Turley, RN, PhD[1]
Kathy A. Johnson-Throop, PhD[1]
Christoph Eick, PhD[2]
Mark S. Tuttle, FACMI[3]

[1]University of Texas Health Science Center at Houston, School of Health Information Sciences, [2]University of Houston, Department of Computer Science, [3]Apelon, Inc.

APPROVAL SHEET

[INSERT APPROVAL SHEET]

DEDICATION

I dedicate this dissertation to the ‘Miller Boys’ that I love so much: Jerry, Todd, Jacob, and Caleb. Their constant support and admiration have made this journey a pleasant one. All of the Miller Boys have enhanced for me the experience of this training beyond both measure and description. While sometimes not welcome (or even enjoyable!), the distractions they created gave me space from this project and time for ideas to ferment. I will always remember fondly Jacob and Caleb’s “participation” in this dissertation (coloring on paper drafts, chewing on books, unplugging computers, and banging the keyboard). I am most grateful to my husband, Jerry, who has been a tremendous support and a motivation throughout all of my years of graduate study, and often understood the significance of my work long before me. Jerry never failed to be enthusiastic, patient, and interested in my research since the day I first mentioned the topic, and has been my most attentive and dedicated listener throughout. While my little guys will probably never appreciate the time and energy that went into this dissertation, Jerry certainly does and will, and shares as much pride in this work as I do. For that I am truly grateful.

ACKNOWLEDGEMENTS

I am deeply indebted to all of my dissertation committee members, who have provided wisdom, support, guidance, and direction throughout this research. I have felt nothing but enthusiasm and support from this group since the conception of this project, and feel honored to be associated with these accomplished individuals: Kathy Johnson-Throop, who pushed me to think critically and explain everything; Dr. Christoph Eick, who provided insight and direction and took great care in reading many manuscript drafts; and Mark Tuttle, who introduced me to the word ‘comparable’ and volunteered many hours to get me to fully appreciate the concept. I am especially grateful to my advisor and mentor, Jim Turley, who is entirely responsible for me considering and entering this doctoral research program. Jim’s creativity and enthusiasm helped shape this project and motivated me to continue. It was Jim Turley that identified my research problem as one of knowledge representation and fusion, and he allowed me to see this is a critical and emerging computational issue in informatics. I am indebted to him as well for his thoughtful, critical, and fast reviews of these and many other papers. I consider myself lucky for his training and for the breadth of knowledge and multiple perspectives that he has given me through all of my years in informatics.

I also would like to thank Dr. Charles Macias and Dr. Marianna Sockrider from Baylor College of Medicine. The time and expertise that both shared throughout this research were critical. Drs. Macias and Sockrider helped me, many times, to determine the relevance of my abstract ideas to their research and clinical practice, and I am grateful to them both. I want to thank the professional and research staff for the Texas Emergency

Department Asthma Surveillance (TEDAS) project, who helped me understand the data, the semantics and the information flow and work processes at all of the institutions involved in this project. I am indebted to many individuals who shared with me their time, expertise, and enthusiasm for research. I especially want to thank Michael DeGuzman, Jennifer Jones, Fahimeh Sasan, Troy Bush, Stephanie Mosley, Balambal Barti, Melinda Tilman, Dr. Ed Brooks, Dr. Stuart Abramson, Taffy MacDoogle, and Judy Keys. In addition, I would like to thank Charles Lindhal and Bryan Barbeau for their genius programming intellect and guidance.

I am indebted to the National Library of Medicine for supporting this research, first as a training grant through the Keck Center for Computational Biology, and secondly as an individual fellowship in applied health informatics. The funding and intellectual support provided by the National Library of Medicine made this dissertation possible, and my attendance at professional conferences over the past 3 years has enhanced my training beyond measure. Within the UT School of Health Information Sciences, several individuals have provided me with support and enthusiasm throughout my graduate studies in health informatics. In particular, I would like to thank former Dean Dr. Doris Ross, Dr. Randy Scott, Dr. Jack Smith, Dr. Constance Johnson, Juliana Brixey, Dana Mejia, and Debbie Todd.

ABSTRACT OF THE DISSERTATION

The current state of health and biomedicine includes an enormity of heterogeneous data ‘silos’, collected for different purposes and represented differently, that are presently impossible to share or analyze in toto. The greatest challenge for large-scale and meaningful analyses of health-related data is to achieve a uniform data representation for data extracted from heterogeneous source representations. Based upon an analysis and categorization of heterogeneities, a process for achieving comparable data content by using a uniform terminological representation is developed. This process addresses the types of representational heterogeneities that commonly arise in healthcare data integration problems. Specifically, this process uses a reference terminology, and associated "maps" to transform heterogeneous data to a standard representation for comparability and secondary use. The capture of quality and precision of the “maps” between local terms and reference terminology concepts enhances the meaning of the aggregated data, empowering end users with better-informed queries for subsequent analyses. A data integration case study in the domain of pediatric asthma illustrates the development and use of a reference terminology for creating comparable data from heterogeneous source representations. The contribution of this research is a generalized process for the integration of data from heterogeneous source representations, and this process can be applied and extended to other problems where heterogeneous data needs to be merged.

TABLE OF CONTENTS

INTRODUCTION TO THE DISSERTATION

The four attached manuscripts collectively represent a PhD dissertation on the integration of heterogeneous health care data. The aim of the entire dissertation was to develop a generalizable process for transforming data with different native database representations into a single uniform representation that would enable comparability of the different data. This comparability is required for any subsequent aggregation, manipulation, communication, and analyses. The hope was and is that this process be applied to current and future problems across multiple domains. Because of these goals, the final process and its description in the attached manuscripts are deliberately abstract and focused on the process requirements rather than the technological implementations.

Title keywords that need to be defined to appreciate this research are 'heterogeneous databases' and 'comparability'. Heterogeneous databases are defined as separate autonomous databases, independently created for unique purposes, with substantial differences in database schema. [1] It is important to recognize that the content of the databases must be considered "semantically" equivalent in a very general sense to be considered a heterogeneous database problem. Semantically similar databases reportedly contain the same "type" of information or constructs. For example, multiple Emergency Department databases with presenting complaint data (however the contents are represented) constitute heterogeneous databases, whereas multiple databases from an emergency room, a laboratory, and a pharmacy would represent disparate (but not necessarily heterogeneous) data sources. Although heterogeneous databases broadly contain the same types of semantic content, the content can be represented in many different ways, often resulting in semantic differences that are difficult to identify and resolve. These different representations across heterogeneous data sources make the data incomparable in their native formats. The data tend to be disparate on two levels: data models and underlying data content. Data models result from different database designs (selected data structures and their inter-relationships). Within each data structure (e.g., attribute or field) there are different knowledge representations (e.g., vocabularies) whose concepts and relationships differ, and whose uses differ depending upon context (e.g., site

of data collection and nature of local coding conventions). The informatics challenges of this research are substantial.

Comparability, a notion that is also central to this research, is a word that many use but few define. Essentially, comparability examines the character or qualities of two or more objects for the purpose of discovering their resemblances or differences. [2] Operationally, we define comparability as the same, or homogeneous, representation of data from multiple sources that permits the determination of equivalency and other relationships (e.g., similarities and differences) between the data. These two definitions (heterogeneous databases and comparability) frame this research. The start state is heterogeneous data representations, the goal state is a homogeneous data representation, and our research result is a process to move from one state to another. Ideally, we would like to arrive at a new, homogeneous data representation that addresses the final information needs, maintains the intended semantics and data granularity of each local source, and captures the similarities and differences across local source representations.

There are three approaches to achieve comparability from heterogeneous data representations. [2] The first most common approach, the *implicit* approach, uses a domain expert or programmer to "recode" instances into a like representation, without explicit rules or logic behind the transformation. For example, data instances of "difficulty breathing" and "SOB" might both be recoded as "respiratory", but the reasoning behind the transformation is not formalized. While there is likely a conceptualization or reference model in the head of the programmer, such implicit reference models cannot be examined, verified, and refined, and they are often at high-risk for the loss of data granularity that results from using a lowest common denominator approach. The second way to achieve comparability, *pair-wise comparisons*, maps equivalent terms from multiple heterogeneous data representations on a 2x2 basis. This approach, while common, is labor-intensive and time-consuming (requires comparisons of each data representation to every other representation in the problem set), and is difficult to scale and maintain. Further, any one of the local models may not be sufficiently expressive to meet final analysis purposes. The third approach to

comparability, the use of a *reference model*, is the preferred approach for integrating heterogeneous data for semantic comparability. The reference model provides one representation to capture all (or all that is "important" from) the local representations, and the ideal structure and content of this referent representation is determined by the final uses of the aggregated data.

A focus of the process generated from this dissertation research is the use of a type of reference model called a reference terminology. A reference terminology is a terminology (i.e., set of specified concepts and inter-relationships) that functions as the standard for comparison of data from heterogeneous representations and/or collected for different purposes. [2] A reference terminology names and organizes concepts relevant to purpose or "use case", and provides the "meaning" of the information-units in the structure. As such, a reference terminology should be understandable, reproducible, and useful. [3]

Two primary prerequisites drove the development of a generalized process for integrating data from heterogeneous representations. The first prerequisite was to understand the nature of the heterogeneities, or differences, across sources representations. An articulated framework for identifying the types of differences that exist between heterogeneous databases was much needed and is a significant contribution of this research. This framework of differences, labeled a framework for understanding representational heterogeneity, is presented in [4], which identifies and characterizes all of the types of representational heterogeneities that exist across multiple database representations, and provides a means to organize current data integration approaches and identify areas in need of future research. The classification of representational heterogeneities typically illustrated across multiple databases was used as a basis for the creation and evaluation of the final generalized process presented in [5].

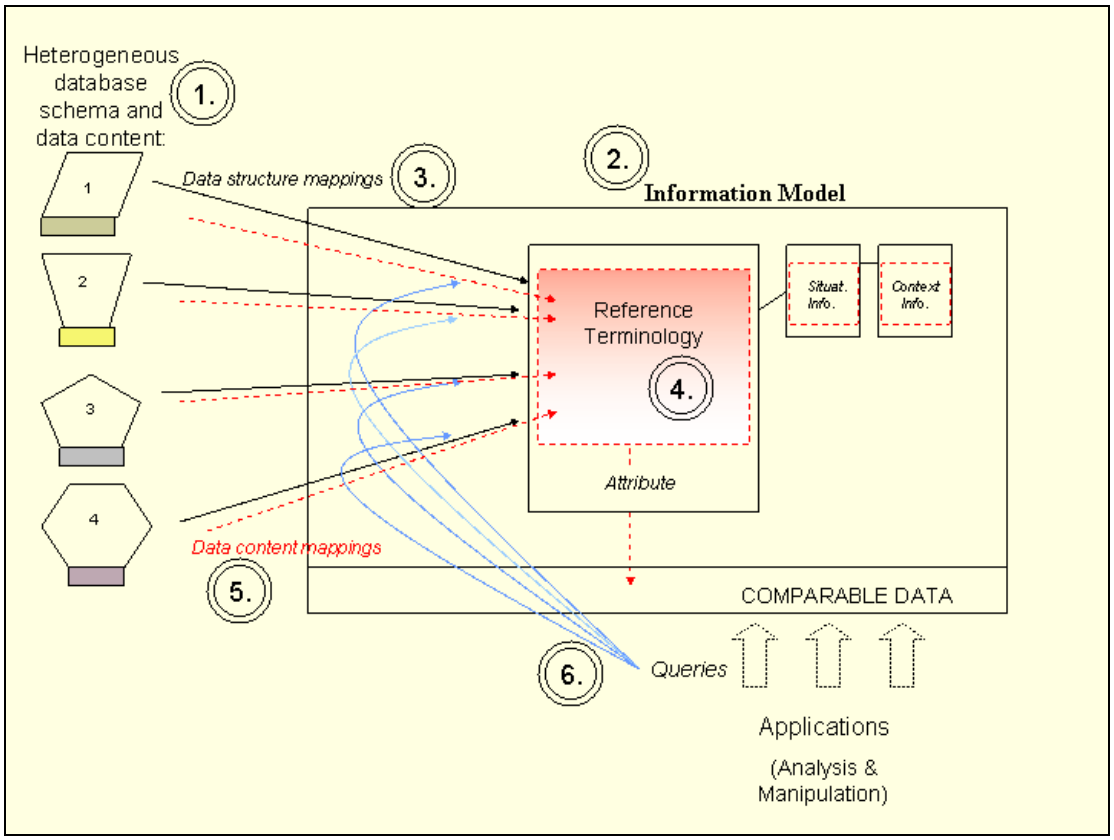
The second prerequisite for the development of this process was to clearly identify the goal of heterogeneous database integration. As mentioned earlier, the goal is comparability (i.e., a homogeneous representation), but there are considerations in the selection and development of an optimal final representation. In [2] we assert that the

goal for integrating heterogeneous databases is to achieve comparable data with a homogeneous representation from different source representations. Further, we define success as the retention of as much granularity (i.e., depth and detail) and intended meaning as possible from each source. To maximize success according to the criteria we have defined, our process places a heavy emphasis on the data-driven or bottom-up development of the reference terminology.

To develop a generalizable process for achieving comparable data, we needed a setting with real data from heterogeneous representations that needed to be integrated. The dissertation research presented here was facilitated by the Texas Emergency Department Asthma Surveillance (TEDAS) project, funded by the Robert Wood Johnson Foundation and managed by investigators from the Baylor College of Medicine Department of Pediatric Emergency Medicine. The data used to develop this process was respiratory-type Presenting Complaints from four Emergency Department (ED) data sets in the Houston metropolitan area. These research results promise one mechanism to assure uniform data sampling across all study hospitals for TEDAS and other clinical research studies. The use of pediatric ED data was a convenient sample to develop this process. However, other health care domains and applications would have functioned as well. Despite the choice of data for the development of the process, the resultant process for creating comparable data from heterogeneous source representations is abstract and generalizable, and can be applied to other data and domains.

The *process* created from this research is discussed throughout the attached manuscripts and is presented in Figure 1. This process was the goal and is itself the result of this research. Essentially, the process transforms data from native representations to final reference model representations. Since by definition a representation is a partial conceptualization, or surrogate of ‘reality’, for a given purpose, the ideal representations are constructed for specific purposes. Therefore, the first step of this process is to clearly define the intended purposes of the integrated data. These defined purposes are critical for the selection or construction of the reference models that homogeneously represent the final transformed data.

Figure 1. Process for Achieving Comparable Data from Heterogeneous Databases



After clearly defining the purpose for which the data is to be aggregated, the first step (#1 in diagram) is to examine and understand the structures and the concepts encoded therein for each local data source, as well as explore context issues that can explain representational differences. This step is critical and involves examining the intended semantic meaning of each data structure. The presence of this step ensures that, as data is later transformed to a homogeneous representation, the intended meaning and the native data granularity is preserved.

The systematic examination of the types of heterogeneities as described in the framework [4] revealed that heterogeneous databases have two very broad levels of heterogeneities – those resulting from different database schema, and those resulting from differences in the underlying data content. These two broad groupings of heterogeneities each became a target for integration processes, namely database schema integration and data content integration. Both require the use of a reference model. The information model (#2 in

diagram) is a referent model that can assimilate different data structures (e.g., data instances, attributes, or tables) into a singular data element. Once the reference information model is selected or constructed (#2), local data structures are mapped, or transformed, to the new structural representation (#3). Heterogeneous data content (i.e., "what is in the fields") is made homogeneous by mapping the local data instances to concepts in a final reference terminology. The development of a reference terminology (#4), and associated mappings (#5 and #6) are addressed in [2]. The semantic focus of our process adds value to current syntactically-based efforts by suggesting a change in focus from purely syntactical solutions toward a semantic-based approach, designed to capture the intended meaning and operational definitions of each data structures. The process that emerged systematically addresses the specific database schema heterogeneities identified in the first framework paper [4], and addresses the importance of representing these differences in the final model to facilitate informed queries (#6) and analysis of the final data.

Using the diagram presented in Figure 1, [5] describes the development of a generalized process for integrating heterogeneous data and addresses outstanding issues in the database schema integration problem. A part of this process is described in more detail in [2], which describes a generalized process using explicit conceptual reference models, called reference terminologies, to resolve the content integration problem. The highly iterative development of a reference terminology is described fully here, including characterizations of changes or iterations in the evolving reference terminology. Together, these two papers provide a blueprint process for the meaningful integration of heterogeneous data from multiple sources to address specified information needs.

The final paper for this dissertation, a results oriented paper, describes the actual implementation and the product of the application of this process to a real health care problem. [6] The final paper confronts the inherent lack of comparability across heterogeneous emergency department data, and describes the application of the processes described above to achieve comparable presenting complaint data where it was

previously impossible. The solution is the development of the Houston Asthma Reference Terminology (HART), and associated "maps", with which locally-coded pediatric ED presenting complaints can be analyzed. The HART solution is empowered by a global data schema for ED visits, which includes explicit representation of native database schema, and quality and precision information that enhance the meaning of the aggregated data, empowering end users with better-informed queries for subsequent analyses. Essentially, this publication describes the process that we implemented to create the comparable data, and what the implementation of that process created in terms of a specific data integration application product.

Together, the four attached papers represent the spectrum of this dissertation research, from the problem definition, literature synthesis, exploration of possible methodologies, to the actual development, implementation, and evaluation of the final generalizable process that was the proposed intent of the research. The four manuscripts that describe the development of this process are targeted to different audiences, namely the Computer Science and Health Informatics communities. Because of the different audiences, and the need for each manuscript to be freestanding, there is overlap in content across the articles, as well as some minor changes in terminology and formatting.

The need to integrate data from multiple, heterogeneous source representations in health care is pressing and growing. The size and complexity of health care delivery and research activities, coupled with the lack of a-priori data representation and storage standards, has created a world of isolated data “silos” that to date cannot be analyzed in aggregate. Currently, the health care domain is overwhelmed with data that is largely incomparable, yet the needs for examining these data are becoming more urgent. Some of the rising costs of health care delivery and experimental drug development could be curtailed by using existing data sources and observational research designs on large populations. Similarly, evidence-based care, which requires monitoring data from multiple sources for long periods of time, could move from vision to reality if comparable data could be obtained across multiple populations and multiple points in the health care system. Issues of patient safety and health care quality are receiving well-

deserved attention and driving efforts to look at aggregate data from multiple sources to monitor health care activities and outcomes. Finally, new attention on bioterrorism detection and population surveillance has drawn the spotlight on the current lack of integration of health care data for public health monitoring. The use of a generalized process to achieve comparable data has enormous potential to positively impact a plethora of health care quality and public health activities across the nation and globally.

Informatics theory and practice as a whole deals with the notion of uniform data representations to overcome lack of terminology and data representation standards in medicine and health care. Indeed, the development, use, and evaluation of many controlled healthcare vocabularies, such as SNOMED, LOINC, and GALEN, represent a large body of informatics research. The UMLS, a major accomplishment and significant contribution of health informatics to address the lack of standardization for health care concept representation, was created to provide a much needed linkage between these different vocabularies. In one sense, the Semantic Network of the UMLS acts in itself as a reference model. The broad nature of the stated purposes of the UMLS however, has shaped a conceptual model that is relatively abstract, which can create a loss of data granularity for many purposes. This process, focused on the semantics and granularity of local data, allows for the development of reference models from the actual data, thereby retaining semantics and granularity that are important to the final intended uses of the combined data.

Inherently, the process created by this research facilitates the use of data for purposes other than that which it was collected. In so doing, we are required to represent the context and quality of the local data that impacts its use it at another level. The need to examine the semantics, or intended meaning, for each concept in the local data is often overlooked in contemporary syntactically-based data integration solutions. The explicit examination of semantics included in this process provides a foundation for a formalized (i.e., machine-readable) examination of semantics, facilitating repeatability, and more importantly, automation. The potential for repeatability and automation is a significant advantage of our process that is not possible of many current data aggregation methods.

The true result, and the re-usable knowledge, of this research is a generalizable and repeatable process that transforms data from different source representations into a homogeneous format in which they can be compared and subsequently combined, aggregated, and integrated in sensible ways. The long-term goal (and the *result* of applying this process) is automation; specifically an understanding of when, how, and to what extent automation is possible. It is evident that the entire process cannot be automated, as human domain and context experts are a critical component of the process. Yet, as informatics moves toward more formalized and sharable conceptual models of health care knowledge, it is likely that parts of the process can be automated, reducing the time and resources required for future data integration endeavors. Future validation of the process in other domains and validation of resultant data representations and their value in applied research areas are future activities for this research.

References

1. Sheth AP, Larson JA. Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. ACM Computing Surveys 1990;22(3):183-236.
2. Richesson RL, Turley JP, Johnson-Throop KA, Eick C, Tuttle MS. Development and Use of a Reference Terminology to Maintain Data Granularity and Semantics in the Integration of Heterogeneous Data Content. Submitted to: Data and Knowledge Engineering, 8-03 2003.
3. Campbell KE, Oliver DE, Spackman KA, Shortliffe EH. Representing Thoughts, Words, and Things in the UMLS. Journal of the American Medical Informatics Association 1998;5:421-431.
4. Richesson RL, Turley JP, Johnson-Throop KA, Eick C, Tuttle MS. Foundations for Heterogeneous Database Integration: A Framework to Identify Representational Heterogeneities. Submitted to: Journal of the Association for Computing Machinery, 8-03 2003.
5. Richesson RL, Turley JP, Johnson-Throop KA, Eick C, Tuttle MS. Creating Homogeneous Data from Heterogeneous Representations: A Process for Heterogeneous Database Integration. Submitted to: Data and Knowledge Engineering, 8-03 2003.
6. Richesson RL, Turley JP, Johnson-Throop KA, Eick C, Sockrider M, Macias CG, et al. Obtaining Comparable Presenting Complaint Data From Heterogeneous Emergency Department Databases. Submitted to: Journal of the American Medical Informatics Association, 8-03 2003.
7. Burgun A, Botti G, Fieschi M, Le Beux P. Issues in the Design of Medical Ontologies Used for Knowedge Sharing. Journal of Medical Systems 2001;25(2):95-108.

8. Sugumaran V, Storey VC. Ontologies for Conceptual Modeling: Their Creation, Use, and Management. Data & Knowledge Engineering 2002;42:251-271.
9. McGuinness DL. Conceptual Modeling for Distributed Ontology Environments. Proceedings of the Eighth International Conference on Conceptual Structures Logical, Linguistic, and Computational Issues (ICCS 2000) 2000(August 14-18, 2000).

A Framework for Representational Heterogeneity*

Rachel L. Richesson, PhD, MPH[1], James P. Turley, RN, PhD[1], Kathy A. Johnson-Throop, PhD[1], Christoph Eick[2], PhD, Mark S. Tuttle, FACMI[3]
[1]University of Texas Health Science Center at Houston, School of Health Information Sciences, [2]University of Houston, Department of Computer Science, [3]Apelon, Inc.

*[SUBMITTED TO THE JOURNAL OF THE ASSOCIATION FOR COMPUTING MACHINERY, AUGUST 2003]

Abstract

Representational differences across heterogeneous databases can arise from several sources, causes numerous types of data conflicts, many semantic in nature, and is the source of the majority of difficulties in the database integration problem. To facilitate the resolution of these heterogeneities, an understanding and characterization of the differences between heterogeneous databases must be defined. A framework for classifying the many representational and semantic differences across heterogeneous databases is presented here. This framework will support the development of tools and processes for which to integrate heterogeneous data while preserving the intended semantics and granularity of the native data sources.

Introduction

The goal for integrating heterogeneous databases is to achieve compiled data with a homogeneous representation from different source representations while preserving native data granularity and semantics. This requires resolution of multiple heterogeneities. Often, these heterogeneities are difficult to identify and require domain expertise to detect. An analysis and organization of the general types of heterogeneities encountered across heterogeneous databases is required to support pragmatic and research activities in this area. This paper provides a classification for the variety of heterogeneities that arise from heterogeneous data sources, and surveys common approaches to overcome them. Heterogeneities encountered across multiple databases are of 3 major types: physical, data model, and representational. The most challenging and outstanding heterogeneous database integration issues are in the identification and resolution of representational heterogeneity and the resultant semantic data conflicts that often arise. This framework presents a classification and description of types of representational heterogeneity by the source (database schema, measurement or concept systems, and context) and by the types of data conflicts that emerge (format, naming, structural, semantic, precision, missing content, and semantic). This framework will support the development and classification of much-needed tools and processes for which to integrate heterogeneous databases in a variety of domains.

Background/Definitions

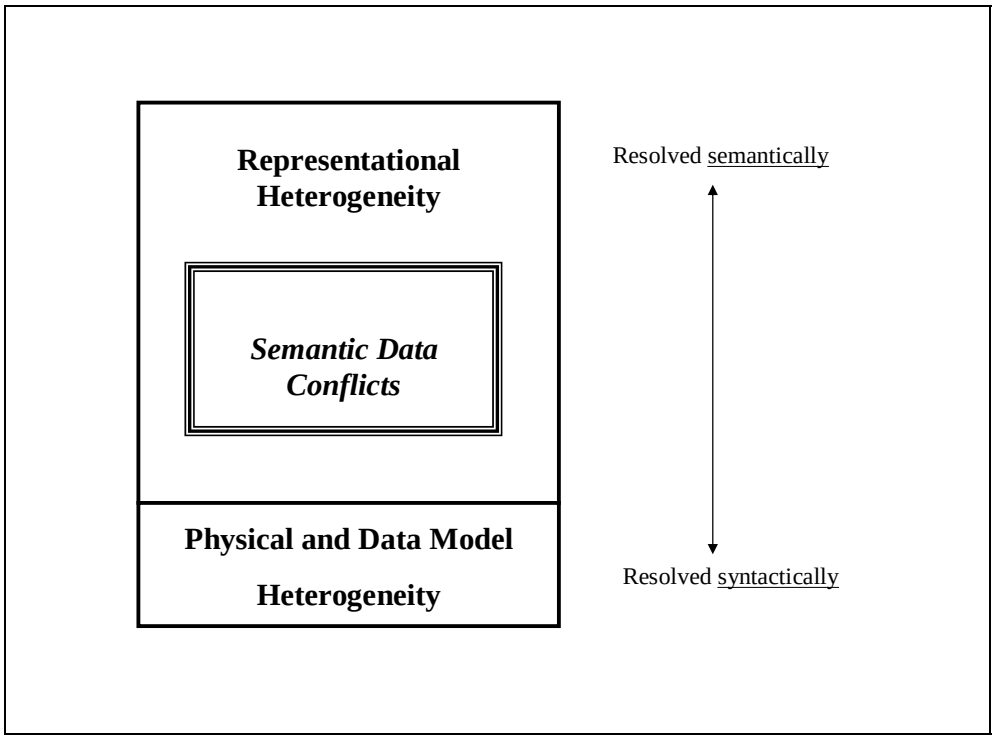
Heterogeneous Databases

Heterogeneous databases can be defined as separate autonomous databases, independently created for unique purposes, with substantial differences in both abstract data models, which represent the underlying paradigm of the database (e.g., flat-file, relational, hierarchical, object-oriented), and database schema, which represent the developer's conceptual model of the data structures and their interrelationships.[1] It is important to recognize that the content of the databases must be considered "semantically" equivalent in a very general sense to be considered a heterogeneous database problem. Semantically similar databases reportedly contain the same "type" of information or constructs. For example, multiple Emergency Room databases with presenting complaint data (however the contents are represented) constitute heterogeneous databases, whereas multiple databases from an emergency room, a laboratory, and a pharmacy would represent disparate (but not necessarily heterogeneous) data sources. Although heterogeneous databases broadly contain the same types of semantic content, this content can be represented in many different ways (broadly termed representational heterogeneity), often resulting in semantic differences that are difficult to identify and resolve.

Types of Heterogeneity

The challenge of creating integrated data with a uniform, or homogeneous, representation from heterogeneous databases is in identifying and resolving all the heterogeneities, or differences, that exist between the source databases. Heterogeneity from multiple databases can be attributed to physical or platform-dependent sources, differences in data models and representational differences that manifest in a variety of data conflicts, many semantic in nature. (Figure 1)

Figure 1. Types of Heterogeneity Encountered in Heterogeneous Databases



Representational heterogeneity and subsequent semantic data conflicts account for the majority of difficulties and outstanding research issues in the integration of heterogeneous databases, and therefore are the primary focus of this discussion. After a brief mention of physical and data model sources of heterogeneity, common representational and semantic differences are characterized and explored.

Physical and Data Model Heterogeneity

The most basic of heterogeneities are those that affect the physical communication of multiple systems. These disparities are related to the hardware or system features (e.g., instruction format, data formats and representation, configuration) or the operating system (file systems and operations, naming of files and file types, transaction support, inter-process communication). The notion of 'data independence' in modern database design (meaning that the data records should remain independent from the application) has largely eliminated these types of compatibility issues, especially as databases and database management systems have evolved over the past 20 years.

Differences in (abstract) data models (e.g., flat-file, relational, hierarchical, object-oriented) from heterogeneous sources consist of disparities in data structures, constraints on their interrelationships, and differences in both capabilities and format of the query languages used to access data in each database. The relational database model is the dominant abstract data model in many industries, including health care, and therefore it is the assumed data model for this discussion. Overcoming disparities between object-oriented and relational database models is an area of current research.[7] However, the

development of network communication protocols and tools, e.g., JDBC, ODBC, DCOM, has made the resolution of data model differences manageable. The greater challenges for achieving homogeneous data from heterogeneous sources are largely semantic in nature. For broader relevance, the duration of this discussion will focus on the classification, sources, and current approaches for dealing with representational heterogeneity and semantic data conflicts.

Representational Heterogeneity

Representational heterogeneity results from the variety with which similar data are represented in different databases. [8] In this paper, we observe a broad definition of representational heterogeneity that includes representational differences resulting from a developer's database design and the operational implementation of an application that could impact the use of the data at the aggregate level. Our definition therefore includes semantic data conflicts introduced by differences in database schema, measurement and concept systems, and context. These representational heterogeneities collectively represent the greatest challenge for the integration of multiple databases, and often the "self-describing" metadata of each database schema fail to represent enough information to detect or resolve them.

Semantic Data Conflicts

Semantic data conflicts are difficult to precisely define, identify, and classify. [1] Broadly, semantic differences occur when there is a disagreement about meaning, interpretation, or intended use of same or related data, and arises from different data type structures, different definitions or conceptualizations of data attributes, differences in coding precision of the data values across multiple databases [1], or context [9]. Semantic heterogeneity in part refers to the fact that data in different systems may be subject to different interpretations, even when data types, labels, and general schemas are identical. [10] There is common consensus that semantic data conflicts (often termed semantic heterogeneities) are the most problematic aspect of heterogeneous database integration efforts. [8] [10] The following section explores key types and sources of representational heterogeneity, including resultant semantic data conflicts, with examples of each from the health care domain.

Framework to Classify Representational Heterogeneity

Representational heterogeneity is often difficult to detect because of the variety of ways it can be introduced into a heterogeneous database system, and because of the many potential data conflicts, including semantic data conflicts, that it can cause. Representational heterogeneity can be attributed to differences in database schema, measurement or concept systems within data attributes, or the context of data collection. The resulting heterogeneities can manifest in a variety of data conflicts, as depicted in Figure 2.

Figure 2. Characterizations of Representational Heterogeneity

	Source:		
Types of Data Conflicts:	Database Schema	Measurement or Concept Systems Encoding Data Content	Context
Format	X		
Naming	X		
Structural: Metadata Conflicts, Compositional, Organizational	X		
Semantic Data Conflicts	X	X	X
Precision	X	X	X
Missing Content	X		X

As shown in Figure 2, representational heterogeneities can arise from differences in database schema, measurement or concept systems encoding the data content, and the context. Each is described below.

Heterogeneities Arising from Database Schema

The database schema denotes the detailed data structures (i.e., relations and attributes in an abstract relational model) and the relationships between them. The schema represents the developer's design of the knowledge domain, and as a consequence there can be many possible valid variations, as shown in Figure 3. Differences in the representation of data structures across multiple databases, collectively termed *schematic heterogeneity* [10], are not trivial and result in a number of data conflicts, including format, naming, structural, semantic, precision, and content.

Figure 3. Sample Data Instances Using Different Database Schema for Emergency Room Data Capture *(note: data from same 2 patients represented in 3 different ways)*

Emergency Room A

Patient #	Date of Service	Age	Chief Complaint	Acuity
123456	10-24-01	12	Cough/Fever/Malaise	Mild
234567	10-24-01	3	Respiratory Distress	Severe

Emergency Room B

Medical Record #	DOS	Time of Service	Age	Acuity
123456	10-24-01	0300	12	Mild
234567	10-24-01	1400	3	Severe

Medical Record #	Presenting Complaints
123456	Cough
123456	Fever
123456	Malaise
234567	Respiratory Distress

Emergency Room C

Visit #	Social Security #	Date/Time of Service	Date of Birth	Description	Value
888	123456	10-24-01 3:00am	11-1-1991	Presenting Complaints	Cough
888	123456	10-24-01 3:00am	11-1-1991	Presenting Complaints	Fever
888	123456	10-24-01 3:00am	11-1-1991	Presenting Complaints	Malaise
888	123456	10-24-01 3:00am	11-1-1991	Acuity	Mild
999	234567	10-24-01 2:00pm	3-6-2000	Presenting Complaints	Respiratory Distress
999	234567	10-24-01 2:00pm	3-6-2000	Acuity	Severe

The different presentations of data instances shown in Figure 3 all represent valid database designs, and their differences are due to the specific information needs of each organization, and to the developer's conceptualization of these needs. In general, hospital emergency rooms share the same workflow: patients present with one or more self-reported complaints or problems, demographic information is collected, patients are assessed by a nurse for urgency or acuity, and finally they are seen by a physician for diagnosis and treatment. Despite the broad similarity in workflow and data capture needs across emergency rooms, a current lack of standards results in an enormous variety of database implementations. The data instances from Emergency Room A, for example, show all information in one table which includes data attributes for a visit identifier, date of service, age of patient, chief complaint, and acuity. The database schema for Emergency Room B, however, includes a patient medical record number, date of service, time of service and patient age in one table that is related to a separate table with multiple instances of presenting complaint values. The sample data instances from Emergency Room C, present yet another valid database structure, with one table containing a visit number, patient social security number, a combined attribute for both date and time of service, date of birth, and a description attribute (presenting complaints or acuity) with the value in the 'value' attribute. All three of these abridged database schemas contain roughly similar information, but the schematic (representational) heterogeneity shown here can lead to a variety of data conflicts, including semantic, which potentially impact the integrity of the integrated data in a multitude of ways, as described below.

Naming or labeling conflicts can also be seen in Figure 3. Equivalent data structures can have different names across emergency room databases (e.g., "Patient #" vs. "Medical Record #", "Date-of-Birth" vs. "DOB", "Presenting Complaint" vs. "Chief Complaint"). Simple naming differences are straightforward to resolve if the meanings of the attributes are the same. However, semantic differences resulting from differing definitions (e.g., the patient identifier at one hospital is unique to the emergency room versus the hospital, or represents a social security number) are the most insidious and problematic to detect, and result in confounding of meaning, discussed later. *Format differences* due to data types need to be represented uniformly to combine data in a valid way. For example, the date values might be in a variety of date formats or string data types across heterogeneous

database schema, and need to be transformed to a common representation prior to any integration.

Structural differences in data elements between different database schema are also straightforward to detect. Three main types of data conflicts arise from structural differences: metadata, compositional, and organizational. *Metadata* conflicts arise when the same construct is represented at the schema level in one system and at the instance level in another. For example, in Figure 3, Presenting Complaint contents (named "Chief Complaint" in Schema A) are represented as distinct attributes in Emergency Rooms A and B, but in Emergency Room C, they are represented at the instance level (i.e., the "Description" attribute contains instances of "Presenting Complaints" and the "Value" attribute contains the specific presenting complaint data.) *Compositional data conflicts* arise when data is represented in one attribute versus multiple attributes across different database schema. For example in Figure 3, Emergency Room B represents the date and time of visit as two distinct attributes, while Emergency Room C represents one attribute for both concepts. A common occurrence of compositional conflicts is the breakdown of address into many attributes (number, street, city, state, zip) versus one text attribute. *Organizational differences*, for lack of a better term, are caused by different quantities of attributes to represent a given concept. For example, one hospital might capture a single attribute for presenting complaints while another might capture 3 distinct presenting complaint attributes or instances. Such organizational differences might have implications for the data. The patient record at one hospital might have skimpy information (i.e., a single complaint), not because the patient failed to have more presenting complaints, but because the constraints of the system limited the capture of other complaints. These organizational disparities can introduce semantic differences requiring the explicit representation of local database schemas to resolve.

Syntactic, rather than semantic, solutions are often sufficient to overcoming the naming, format, and structural (metadata, compositional, and organizational) conflicts described above. The distinction being that resolutions for syntactic problems are achieved via programming syntax, whereas resolution of semantic differences requires exploration of the context and intended semantic meaning of the original data structures to support any programming solution. For example, it is easy to envision systematic approaches to moving data from instance level to attribute level to table level, with out necessarily invoking a change in meaning. All of the above schema differences manifest in representational heterogeneities; those schematic variations that lead to differences in *meaning* at the aggregated level represent semantic data conflicts

Semantic data conflicts result from variable definitions of the contents of a given data attribute across different database schema. This confounding in meaning can arise when database schema and names are identical, and therefore this often slips by programmers and automated integration processes. Too often, these differences are non-explicit, and perhaps subtle, and require careful investigation to detect. For example, two different emergency room databases might include a data attribute called "Presenting Complaints". One emergency room might collect patient-reported complaints but also routinely include observations from the triage nurse, where another hospital might only record complaints

stated by the patient. This results in semantic differences in the content of what appears to be a similar data attribute across heterogeneous sources. These types of semantic data conflicts, the most difficult manifestations of representational heterogeneity to resolve, require an examination of intended meaning to detect and resolve.

The final type of data conflicts arising from different schema, *content differences*, occur when data represented in one database are not directly represented in another, due to data structures that are implicit or simply missing.[8] Implicit data structures arise when data is obvious, and therefore not represented, in the local context (such as the name of the emergency room) but becomes important when data is being aggregated and examined globally.[9] This type of heterogeneity arises from the context of data collection and is discussed in detail later. Implicit data structures or content can often be derived (e.g., age can be derived from the date of birth and date of visit via simple calculation.) Yet, it is important to note the directionality of these derivations (e.g., age can be derived from date of birth, but date of birth cannot be computed from age.) Finally, content differences arise when an attribute (e.g., "admitted to hospital?") is included in one data schema but not another. The uniform representation of the combined data should distinguish people whose discharge status was not collected (i.e., missing at the attribute level) from those whose hospital admission status is missing at the instance level, as semantically the missing data has different meanings. Such disparities in content reinforce the importance and need for further research regarding the representation of "missing" data.[8]

Typically, database schema do not describe data attribute contents beyond data types, and the specifications of database schema often do not contain enough information to resolve many representational heterogeneities and semantic data conflicts. [1] Exploring the content of the data attributes opens up more problematic semantic differences, including heterogeneities from disparate measurement and concept systems encoding the data content, as discussed in the next section.

Heterogeneity Arising from Different Measurement and Concept Systems

A common data integrity challenge for heterogeneous database integration efforts is the assurance that the measurement systems are *comparable* across attributes that need to be combined. Comparability is a broader notion than equivalence, and implies the need for a common representation to make judgments of relationships between the values (e.g., equivalent to, greater than, less than; broader than, narrower than, etc.). Classic examples of measurement differences include length in feet vs. inches, or weight in pounds vs. kilograms. Since the conversions for these different ratio systems are well known, these disparities reduce to a common problem of scaling, resolved by simple re-coding to a standard measurement system. However, resolving disparate measurement systems involving nominal or ordinal data are more challenging.

Many nominal and ordinal coding schemes, including vocabularies and terminologies, are systems of concepts. Assimilating different concept systems can result in potentially serious precision and semantic data conflicts. Consider the acuity (a measure of the

seriousness of patient's condition) measurement systems differ across 3 emergency rooms as shown in Figure 4. *Precision conflicts* arise when the units of measurement are not comparable for similar structures across heterogeneous databases. It is clear from Figure 4 that each of the 3 emergency rooms uses a different coding scheme to represent acuity information, and that the granularity of these scales differ. Two of the measurement systems (Emergency Rooms A and C) represent acuity on a 3-value ordinal scale, while the other (Emergency Room B) represents this same construct on a 4-value ordinal scale. Even without an understanding what concepts the specific values represent, it is apparent that no combination of these coding systems will result in a singular system that represents the granularity of all of the local codes. Mapping to a single component coding system implies either the loss of data granularity from some coding systems, or the need to impute imprecise concept mappings from others.

Figure 4. Alternative Coding Systems for Acuity Information

EMERGENCY ROOM A	EMERGENCY ROOM B	EMERGENCY ROOM C
ASAP	Red	Team
Urgent	Blue	Check
Stable	Yellow	Shock
	Green	

Semantic data conflicts arise when the concepts are not comparable for similar structures across databases, and this lack of comparability is often difficult to detect. It might seem logical to assimilate the 3-value scales for Emergency Rooms A and C, but if the code values "ASAP", "Urgent" and "Stable" do not represent the same underlying concepts as "Team", "Check", and "Shock", this would lead to a semantic mis-match, or confounding of meaning, in the aggregated data.

This simple example is typical of problems encountered in controlled healthcare vocabularies with hundreds of thousands of concepts. The enormity and complexity of medical knowledge makes the assimilation of different vocabularies a greater challenge than dealing with many conventional concept and measurement systems. Because the concepts represented in heterogeneous medical concept systems are often not 1:1, or even n:1, they are very difficult to resolve. The translation between disparate units in ratio measurement is straightforward, since, for example, one inch always equals approximately 2.5 cm. But what is the relationship between "coughing/wheezing" and "breathing problems", or the relationship between "nasal congestion" and "runny nose/ green"? The existence of many concept and measurement systems for health care knowledge is frequently referred to as the "vocabulary problem" in the medical informatics literature, and their resolution is a major research focus for the field. [11] [12] Further, multiple conceptualizations and representations for temporal data also challenge data integrity efforts, and a clear understanding, uniform representation, and explicit distinction between "database time" and "event time" should be captured in the integrated data. [13]

The Role of Context

The notion of context has been conceptualized in many different ways, and is relevant to the identification and resolution of every type of representational heterogeneity (including semantic data conflicts) described this far. First we will describe different operational definitions of context. Secondly, we will describe the role of context in detecting semantic data conflicts, and identify key elements of context that can resolve the data conflicts caused by representational heterogeneity across multiple databases.

At the most basic level, context denotes the symbols or characters surrounding a term or underlying concept of interest, usually within a data value. This type of context is focal to natural language processing, information retrieval, and many web-based search applications. [14-17] Mathematical and computational algorithms measure frequency and relationships of words or concepts to calculate real-world distances, semantic distances, establish domain context, or establish importance. In unstructured environments such as the Web, context generally is defined in this way and is used as a measure for retrieving and measuring the quality (i.e., relevance) for matching Web-based resources. [18] Identifying context as this level has implications for heterogeneous database integration, and is highly relevant to determining equivalencies in unstructured (i.e., free-text) data attributes within and across multiple databases. The assimilation of unstructured free text codes, for example, might employ natural language processing techniques to identify that “cough” and “no cough” are not equivalent, while “cough” and “coughing” are.

In the area of database integration, however, context generally refers to the native database schema, specifically the relationships between data attributes. [19] The content of one data attribute can influence the implied content of other data attributes, and therefore it is often necessary to consider an entire local data schema in data integration efforts. [19] For example, an attribute name “Family History of Disease” in one database schema is semantically equivalent to a combination of attributes “Condition”, “Pertains to”, “Family Member”, “Temporal Marker” (= past) in another database schema. Further, different data types and constraints between data structures across heterogeneous database schema can impact the semantic understanding of the data at the global or aggregate level. The distinction between free-text (string) versus coded data content might be important in assimilating data from presenting complaints across multiple schema. A free text entry of “cough/secretions” likely means that both of the concepts “cough” and “secretions” are present with certainty, whereas in a coded data attribute, the value “cough/secretions” could mean that the coder intended either “cough” or “secretions” or both. The database schema context is also important to distinguish between patient records with one presenting complaint concept due to constraints in the local database schema (i.e., the schema only allows one presenting complaint to be entered) versus the reality of the clinical situation (i.e., the patient truly had only one presenting complaint). The semantic understanding and resolution of these issues is driven by the purpose of the data integration, but also by the explicit contextual representation of each local database schema.

A broader perspective of context, and one most critical to resolving semantic heterogeneity, relates to organizational and process issues that influence data collection and impact the semantic meaning of the data. Many database heterogeneities are due to such organizational or data collection contexts. Dampney et al. (2001) note that implicit or even obvious information at the database level is often not represented when using the data in another context. For example, clinical information systems in emergency rooms are designed to capture data related to organizational functions and clinical care. These data models do not explicitly code data that are implied or unnecessary for the database's intended purpose (e.g., all patients in an emergency room database were observed in the emergency room, all physicians at a children's hospital are pediatricians, presenting complaints are selected from a locally created symptom list, the use of a beta-agonist for an asthma patient implies that the medication was inhaled via a nebulizer). In health care, implicit information structures that are not represented include context of role, organization, purpose, and data classification schemes. [9] Context of data collection is unique to each data source and each data observation. To leverage the data from these databases for other research purposes, relevant information structures that are implicit in the context of the original data source (e.g., the setting of patient care or the granularity of knowledge representation deriving from the classification schemes used) must be identified.

This conceptualization of context can be extended to organizational or other factors that impact data collection. For example, the general model of emergency care is that patients arrive, state a complaint or ailment, are triaged by a nurse, and then diagnosed and treated by a physician. Although usually not represented in the database, it is understood in the health care domain that a "presenting complaint" is reported by the patient. Therefore, a presenting complaint of asthma (reported by patient) carries a different meaning than a diagnosis of asthma (reported by a physician). Even more problematic are cultural or sociological climates that influence the data definitions in subtle ways. For example, one hospital might pressure its ER physicians to give an asthmatic child a diagnosis of asthma, while another hospital might feel that a diagnosis of a chronic condition, such as asthma, is not appropriate in an acute care ER setting, or that the diagnosis in children is a potential source of stigma and anxiety, and thereby pressure physicians to record a less specific diagnosis, such as Reactive Airway Disease. This variety in data coding procedures is rarely evident in a database schema and can bring to question the comparability of the integrated data.

General Strategies for Database Integration

The representational heterogeneities described above result from differences in database schema, underlying measurement and concept systems, and context. Representational heterogeneities across multiple databases lead to many types of data conflicts, including format, naming, structural, semantic, precision, and content. It is important to acknowledge that 1.) it is inherently difficult to tease out all of the different sources of heterogeneity, 2.) that a given source of heterogeneity might manifest in multiple data conflicts, and 3.) that many solutions address multiple heterogeneity problems. The next

section describes basic database integration approaches and requirements, and identifies the representational heterogeneities and resultant semantic conflicts that they address.

The Purpose Is the Driver

Any strategy for database integration to create a homogeneous data representation must include a thorough analysis of the process and informational needs. [20] Of practical importance are the process needs of access updates and ownership and control. The logistics of data acquisition are a primary concern, and questions about the process needs (including frequency and scalability) should be addressed. These issues are discussed extensively elsewhere. [1, 21]

General Approaches

Two broad approaches are used to achieve homogeneous or comparable data from heterogeneous databases. One strategy is to extract the desired data using query language specific to each data base, and then "translating" the data from each source to a uniform representation to achieve comparability. This data-translation strategy is common in data warehouse or clinical data repository projects. [8] The aggregated comparable data is accessed by the user using the query model of the final data repository, and users are oblivious to any representational differences across component databases. This data-translation process, also called *data integration and summarization*, requires periodic data export and integration from each data source. The second approach involves "translating" a desired query into equivalent functional queries for each local data source to extract comparable data from each source. Most strategies for query translation, also called *query modification*, involve information mediators or "wrappers" for each local information system that describe what the databases can provide in terms of the local abstract data model and database schema, and what types of queries they can answer in terms of the native query language. [18] This query- translation approach is difficult to implement but can provide more timely access or real-time data. [8] Various models of database federation can be considered to determine which strategy best suits the project requirements in terms of access, control, and availability of updates. [1] Both data-translation and query-modification strategies, however, ultimately require the same thorough examination of each database schema, underlying measurement and concept systems, and context to ensure the validity of the compiled data. This thorough examination is through the looking glass of a broader conceptual model of the domain and an understanding of the purposes driving the data integration effort. The definitive goal of this examination of heterogeneities is to achieve semantic comparability (a uniform representation and a consensually understood *meaning*) of data from heterogeneous source representations.

The Ultimate Goal Is Comparability

Whether the approach is data-translation or query-modification, the challenge of integrating heterogeneous databases is to make the information comparable on all levels. The databases are comparable if the semantic intent of the data can be transformed to a homogeneous representation. This uniform representation is considered "global" relative

to the local component data sources, and requires choices to be made about the platform, abstract data model, database schema, and most importantly the measurement system "units" or precision for each data attribute. Once these choices are made, the transformation process requires mappings or defined relationships, between both data structures and underlying content, to these defined uniform knowledge representations. In the case of heterogeneous data models and schema, the defined uniform representation is called a *global* or *reference schema*. In the case of concept systems encoding data attribute content, the defined uniform representation is called a *reference terminology*. Both global reference schema and reference terminologies are conceptually-based referent standards that together form the uniform representation for the aggregated data. These conceptually-based referent standards can be created anew or by integration of the underlying data sources. The comparability that they enable may entail loss of data granularity or precision from some local sources; the most successful comparability solution is that which preserves the semantic intent and the most data granularity from the most sources.

The Key is Conceptually-Based Reference Standards

Resolving Representational and Semantic Heterogeneity

The variability in database schema and measurement systems, as well as differences in context at multiple levels, create enormous potential for loss of meaning when aggregating databases in any domain. To overcome these representational and semantic heterogeneities, the semantic intent should be the focal point of data integration efforts, and therefore conceptually-based knowledge representations, at several levels, are critical. Successful solutions for preserving the intended meaning of data require the use of one or more conceptually-driven global reference models, which form the blueprint for identifying, understanding, comparing, and ultimately resolving semantic differences from multiple sources. These conceptually-based referents provide the structure for the uniform representation to which heterogeneous representations are transformed. In the attribute of heterogeneous database integration, these conceptual models generally fall into two categories: global database schema that address schematic and context disparities, and reference terminologies that address disparities in concept or measurement systems.

Resolving Heterogeneities from Database Schema

Most processes for heterogeneous database integration involve some type of transformation of local database schema, often to a master or global schema. This global schema defines all of the important data structures and relationships required at the aggregate level, and as such, it forms the limits of what the new data attributes and relationships can express. The global schema guides the query of the integrated data (in the data-translation approach) or defines the translations, or mapping, of local heterogeneous query models to the referent in the query-modification approach. The global schema can be thought of as the schema of a final integrated database, or the "ideal" schema in terms of the final purpose. Global schema can also be thought of as

information models that represent how the data attribute “units” can be assembled into meaningful (patient) records. [22]

In the case of integrating heterogeneous Emergency Room databases, a global schema could consist of constructs and relationships that are common to a general view of Emergency Room visit processes: patient information, visit information, presenting complaints, symptom data, diagnosis, and discharge information. Relevant local structures would be mapped to the global schema to achieve a uniform representation. This mapping can be used to overcome data conflicts caused by schematic heterogeneity, specifically naming, format, structural, confounding, and content conflicts. For example, local data structures such as “chief complaint”, “presenting complaint” would be mapped to the master “Presenting Complaints” attribute in the global schema, yielding semantically-like data in a homogeneous structural representation, thereby eliminating naming differences. The global schema, as any other database schema, includes attribute definitions, formats, and relationships between data structures, which guide the valid mapping of local data structures to global data structures, addressing many of the representational disparities identified in the framework presented earlier. The global schema also has the potential to address semantic data conflicts, if the operational data definitions are carefully and systematically investigated across local sources. Since these operational data definitions are often not explicit, experts and local users are needed to determine the intended and actual semantics of each structure. A global schema does not do the work, but does provide a blueprint of what data definitions and relationships to investigate.

Ontologies can serve as a blueprint for the construction of task appropriate global schema. The conceptualization of a global schema as a representation of general domain constructs and relationships, matches the accepted definition for ontology. An ontology can be thought of as a conceptualization of domain knowledge. As a conceptualization, or knowledge representation, ontologies inherently provide a limited view or surrogate of important concepts for a given purpose, and facilitate computational applications. [23] The important role of ontologies as global conceptualizations of domain knowledge in the integration of heterogeneous data sources has been identified. [18, 24-27] In schema integration, the ontology serves as the global (referent) data schema, and labels (including synonyms) from each of the data attributes from the local database schema can be matched to the ontology to determine which terms are common to both. [28]

A global schema can be created from two different approaches: a bottom-up *schema integration* (literally combining schema from existing heterogeneous databases), or a top-down *schema creation* (driven by a broader conceptual organization or purpose-driven view). While both approaches have their merits, the construction of a global schema is often highly iterative and therefore involves some element of both approaches. Strategies for the design, adaptation, and integration of domain ontologies provide good resource for design of global schema. [24, 25, 28-31] Even in a top-down approach, the design is impacted by the component data elements it needs to capture. A bottom-up approach of integrating ontologies will yield a very different result than top-down methods[28, 31], and the same holds true for schema integration versus schema creation. The former data-

driven approach can have greater potential for automation, whereas the top-down global schema creation strategy begins conceptually by identifying important constructs and concepts to model, usually identified by domain experts. Ultimately, the data integration purpose and needs dictate the optimal approach. However, for resolving semantic disparities related to context, described later, the top-down approach has more value and more extensibility.

Some have argued against the scalability of a global schema approach, claiming that a global schema is too broad in scope to maintain, and that the updating of mappings from the local to the global schemas required for every local data schema change is labor intensive [32] and assumes too much domain knowledge on the part of the end user. [33] To address the difficult maintenance of global ontologies, some advocate only the maintenance of selected linkages based upon relationships between certain (most important) terms in multiple ontologies. [32] To reduce the burden, abbreviations have been proposed where comparisons be implemented on a one-to-one basis, and only with those parts of the schema needed in the final application. There is some merit to all of these arguments, and the scope of the global schema practically should be limited to important content for the intended application. Yet, we contend that a uniform conceptually-based representation in the form of a global database schema or ontology is required for any kind of sensible integration. The feasibility and scalability arguments reinforce the case for tools and methods to construct global schemas of varying scopes.

While the use of a global schema is integral to resolving differences in database content, and context (to be elaborated later), this approach fails to address representational differences resulting from different measurement and concept systems encoding the underlying data. Here, similar arguments apply for the use of other conceptually-based reference standards, called reference terminologies, to uniformly represent disparate measurement systems.

Resolving Heterogeneity from Measurement and Concept Systems: Reference Terminologies

Issues surrounding differences in measurement and concept systems, and the semantic confounding and precision conflicts that result, can be viewed as a microcosm of the issues for global schema integration described above. To define how heterogeneous data structures can be validly aggregated or compared, content from heterogeneous sources needs to be represented within the context of unifying referent standard or conceptual knowledge representation framework. In the case of concept measurement systems, this knowledge representation is called a reference terminology. A reference terminology is a terminology (i.e., set of specified concepts and inter-relationships) that functions as the standard for comparison of data from heterogeneous representations and/or collected for different purposes." A reference terminology names and organizes concepts relevant to the purpose or “use case” and provides the meaning of information units in the structure. [34] A reference terminology for presenting complaints, for example, would identify the relationships between concepts such as “cough”, “wet cough”, “nasal congestion” and “wheezing”. In medical knowledge representation, these relationships are often

represented in hierarchical organizations of concepts. Regardless of the format or complexity, a reference terminology identifies and names the concepts of relevance to the application in a clear, non-ambiguous, and non-redundant way. [34] Mapping is defined as the relation between the representation of a concept in one terminological system to the most similar representation in another system. [35] Reference terminologies and associated mappings can resolve the semantic and precision data conflicts, when the local data values are mapped to the reference terminology in a way that preserves their intended semantic meaning.

Medical informatics has many controlled (standardized) vocabularies that are potential reference terminologies for data integration efforts, but each is better suited for some purposes than others. [36] It has been stated that for any given purpose, no existing terminology will suffice. [37] Operationally, a reference terminology should be understandable, reproducible, and useful. [38] There is little written about methodologies for creating reference terminologies, but accepted standards for good terminologies do exist. [39] [36] Like global schema or ontologies, reference terminologies can be conceptually-driven or data-driven (top-down vs. bottom-up), and the two approaches can create very different representations. Often a top-down approach is dictated by the project purpose. This approach starts with the identification and organization of important concepts for the “use case”. A theoretical or purpose-driven conceptualization drives the content and organizational structure of a reference standard in the top-down approach. A bottom-up approach is data-driven. The content from all local measurement systems or representations is examined in context of each other, and resolution is achieved by trying to merge the data into a common, or homogeneous, representation. However, many of the concepts represented in disparate coding systems are not one-to-one, and many coded terms represent multiple or “lumped” concepts that can influence the reference terminology structure in a bottom-up or data-driven approach. To illustrate, two alternate reference terminologies are presented in Figure 5. Each has implications both for how the data can ultimately be used, and for the precision of the mapping of terms from local data representations.

Figure 5. Potential Reference Terminology Representations and Sample Data Instances

Representation #1:	Representation #2:
Symptoms	Symptoms
Fever/Infection	Possible Infection
Fever	Fever
Infection	Non-Febrile Evidence of Infection
Cough/Secretions	Cough
Cough	Dry Cough
Secretions	Wet Cough
	Chest Secretions
	Nasal Secretions

Precision of mapping of data instances is affected by Reference Terminology structure:
Data Instance: "Cough"
Data Instance: "Cough/Secretions"

The structure the reference terminology is ultimately left to the designer, but not without implication. The question of whether lumped concepts such as "Cough/Secretions" or "Fever/Infection" necessitate a similar grouping in the reference terminology, or whether the driving needs necessitate representing these concepts separately, is very important, and illustrates the implications between a bottom-up data-driven vs. a top-down design approach. Given the two possible reference terminology representations in Figure 5, it is clear to see that the precision of mapping of local data instances such as "cough" and "cough/secretions" can have different levels of confidence.

A uniform knowledge representation is necessary to achieve the comparability required for the meaningful compilation of data content from heterogeneous database representations. The alternative to a single reference terminology is to map each component terminology or concept system to every other terminology for all databases. Such mappings would allow data content from any or all component databases to be "viewed" within the conceptual framework of any one database's concept system. While this would lend more information, the approach is much more laborious, less likely to scale, and more difficult to adapt to changes across the component terminologies. Also, such an approach is limited to the content and structure of the component concept systems and not likely to address new semantic representations that might be required for the secondary analyses, or re-use, of existing data.

While classic approaches to heterogeneous database integration view "what's in the attributes" as a coding problem, the medical informatics research community identifies this as a *knowledge representation* problem, requiring an examination of the underlying semantics of each coded value. As such, the resolution of these terms depends upon conceptualizing each code value not as a "term" but as a (terminological) representation for a unique concept. The use of a reference terminology to resolve differences at the data content level can be likened to the use of a global schema or ontology to resolve

differences at the structural level. Indeed, a detailed ontology could be used for this purpose. The difference between a reference terminology and an ontology can be subtle and is usually a matter of degree. Both can be used as a reference standard for assimilating disparate concept systems, but reference terminologies tend to be more purpose-driven, and less elaborate in terms of formal descriptive characteristics.

Using Context to Identify and Preserve Semantic Intent

The strongest tool for the resolution of heterogeneous data representations is an explicit representation of context. The lack of consensus on a definition or representation of context makes this an ongoing research challenge. The context of individual terms at the data value level is critical to the selection or design of a reference terminology to resolve heterogeneities in units of measurement or concept representation systems. Additionally, the source of data content can also be considered a type of context. In healthcare, the role (patient, nurse, physician) of the person reporting and entering each piece of data affects the value of the information at the aggregate level. The sources of the data are also important. For example, the National Library of Medicine's MEDLINE document management system recognizes that the source of information is a measure of quality that impacts the user's decision-making, and labels publications by type (e.g., peer-reviewed journal article, conference proceedings, on-line source).

Most context-related issues can be identified and resolved with a guiding conceptual model or global schema. The global schema identifies what is relevant to specify in the domain, and thereby guides integrators to "fill in the holes." This is particularly helpful in identifying the implied concepts from data collection or organizational issues. Domain-specific data collection process models facilitate the imputing of context items (e.g., reported-by, logical and temporal sequence of data, etc.) that need to be imputed. Similarly, the global schema approach can resolve differences at the record level. Since the global schema is essentially a new representation of the local database schema, it is the limiting factor for processing of that data, and therefore all important concepts must be contained. For many implied concepts, a top-down schema creation might apply. Unless the notion of who reported the disease is explicitly in one of the databases, the schema integration approach might miss it. If the "reported by" or "hospital name" constructs are important at the global level (or present in some data bases as in schema integration approach) then this blank "attribute" would prompt integrators to determine the information and decide how to impute it. Often, a broader and more detailed ontological domain view, if available, can identify differences in context across sources that might be missed in a data-driven global schema integration approach. The selection or development of the guiding ontology should be driven by the important information needs of the aggregate data.

Identifying differences in data from context due to implied concepts or data collection procedures requires domain experts and persons familiar with the context of data collection. Since the context might differ for each data source, information to resolve disparities must be obtained from representatives familiar with both the conceptual database design and the routine coding procedures for each data source. This often

involves imputing or deriving implied data structures and content for some or all data sources. The gathering of this information might require many research techniques (structured interview, focus groups, observation, etc.) routinely used in other attributes, such as psychology, sociology, and education, and is guided by using a conceptual model for the integrated data that identifies important concepts and semantic relations. In addition, domain experts familiar with the broader domain knowledge and process models will be necessary to identify explicit representations for the disparities in organizational and data collection context within the broader framework and needs of the data integration effort.

Context is particularly sensitive to setting and its resolution will ultimately require developers to observe or query people in the data collection setting. The roles and training of persons collecting and entering the data at each local source should be compared, and possible sources of disparity in data definitions and data collection procedures should be explored. The conceptual model of the domain that is required to represent important similarities and differences between the semantics of component data structures and content should extend to include relevant aspects of context semantics. The focus for both is on the *meaning* of each concept. Domain experts are critical to assimilate such information into the uniform representation. Domain experts are also necessary to form meaningful representations for the inherent variability in the mapping relationships between local terms and concepts in the global schema or reference terminologies.

The Hitch: Quality

The mapping of heterogeneous data representations (whether at the data structure level or the data instance level) to a uniform representation implies that the nature of these mappings will differ by source. The variability in precision of these mappings can be considered a measure of quality. Making this variability explicit can improve the utility of the integrated data and can have implications for the maintenance of the reference standards.

Several domains have developed explicit representations for the quality of mapping to referent standards, but there is no common representation in field of database integration. The ISO defines different types of relations between terms and concepts when transforming between knowledge representation systems, including synonomy, quasi-synonomy, antonymy, monosemy, polysemy, homonymy. [35] In the field of Information Retrieval, a variety of quantitative techniques allow the user to determine the relevance of the returned documents or records. [18] The National Library of Medicine's Medical Subject Headings (MeSH) cataloging system uses a 'broader than' / 'narrower than' classification to characterize matches in the searching of published and indexed medical literature. Applications in statistical linguistics use data-driven quality matches, noting the number and location of important words and word distances in the source document as a measure of quality. [15]

The explosion of information on the internet has created new challenges to aggregate heterogeneous data sources, and many of these applications incorporate the notion of match quality. Many applications, including the COINS project, perform matching algorithms on the text and its immediate context. [18] Another more sophisticated application, designed to facilitate the searching and integration of ontological databases on the Web, allows users to identify concepts from one or many ontologies and specify constraints between them, including: pre-requisite, mutually inclusive, mutually exclusive, and temporal. [28] The LARKS [18] system, another ontological-based matchmaking application, identifies 3 types of matches between data instances: 1.) "Exact" - when both descriptions are equivalent (either literally or by synonymy, or equal logically by logical inference, 2.) "Plug-In match" - when one description can be "plugged in" to another (e.g., broader, narrower, part-of), or 3.) "Relaxed match" – uses a numerical distance value to determine the closeness of two descriptions. These types of approaches could be applied to either the mapping of data values to a reference terminology or data structures (e.g., attributes) in the assimilation of heterogeneous database schema. An enumeration of these types of quality matches between data structures or instances to a uniform representation has potential to facilitate automated mapping strategies and could increase the final data quality in heterogeneous database integration efforts. Optimal definitions and representation of quality will vary by application. Currently,, standard representations for context and quality of maps to conceptual referents are lacking, leaving the integrator and domain experts to make judgments in the translation of heterogeneous databases to a uniform representation. Future research in the development of these explicit representations is critical to advance automated processes to achieve valid homogeneous data.

Conclusions

The most challenging and outstanding heterogeneous database integration issues are in the identification and resolution of representational heterogeneity and the semantic data conflicts that often arise. This framework presents a classification and description of types of representational heterogeneity by the source (database schema, measurement or concept systems, and context) and by the types of data conflicts that emerge (format, naming, structural, semantic, precision, missing content, and semantic). This framework will support the development and classification of current and future tools and processes for which to integrate heterogeneous databases in a variety of domains.

All approaches for heterogeneous database integration share the same broad goal of presenting useful and comparable integrated data, allowing the user to focus on their tasks, rather than the different representations or interpretations of local systems or the conversions thereof. Various representational heterogeneities have been identified in the framework presented here, many of which require resolution of domain and context-dependent semantics. While the process of validly integrating data from heterogeneous databases into a common representation is highly dependent upon the domain, purpose, and local contexts, the overall goals and measures of success are similar. Data from heterogeneous databases cannot be validly aggregated or compared without a uniform or homogeneous representation. The ultimate goal of heterogeneous database integration is

to create “comparable” data in a representation suitable for a given purpose. Given the final information needs, the most successful heterogeneous data integration strategy should retain as much granularity from as many sources as possible.

Successful solutions for preserving the intended meaning of data require the use of one or more conceptually-driven global reference models, which form the blueprint for identifying, understanding, comparing, and ultimately resolving semantic differences from multiple sources. The classic teams of technical programming or database experts must be expanded to include interdisciplinary teams of local database users and domain experts that can identify and preserve semantic intent while transforming data to standard representations. The integrity of the transformed data can be enhanced by including explicit representations of context and quality, and further research into these representations is critical to the success of future endeavors.

Future research is needed in describing methods to develop global schema, including the design of standard ontologies for global schema. Critical to this is the development of explicit representations of context, including representations that would help determine implied and missing concepts. Similarly, methods and tools are needed to develop data-driven and purpose-driven reference terminologies and to evaluate quality of existing ones. Quality characterizations of the mappings between local data structures and uniform representations are warranted. Strategies to delineate operational definitions for data attributes are also needed. Explicit representation of these operational definitions will facilitate future computational methods for aggregating data from heterogeneous representations to a uniform homogeneous representation. Methods for mapping to referent standards should be researched across domains. For all of the above, the emphasis should be on repeatable or reusable methods across domains and to represent and transform rapidly changing local data sources.

The issue of aggregating data across heterogeneous databases is an important problem across all industries and domains, and likely one that will not go away. This framework provides the basis for identifying heterogeneities across multiple databases, and lays the foundation for the classification, development, and evaluation of generalizable processes for heterogeneous database integration.

ACKNOWLEDGEMENTS:
The authors wish to thank Dr. Charles Macias from the Department of Pediatric Emergency Medicine at the Baylor College of Medicine and Dr. Marianna Sockrider from the Department of Pediatric Pulmonology at the Baylor College of Medicine for their expertise, assistance and enthusiasm throughout this project.

This research was facilitated by the Robert Wood Johnson Foundation: Managing Pediatric Asthma: Emergency Department Demonstration Program; Pediatric Texas Emergency Department Asthma Surveillance (TEDAS).

Rachel Richesson is funded by National Library of Medicine Fellowship in Applied Informatics #1 F32 LM07188-01A1.

REFERENCES

1. Sheth, A.P. and J.A. Larson, *Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases.* ACM Computing Surveys, 1990. **22**(3): p. 183-236.
2. Raschid, L. and Y.H. Chang, *Interoperable Query Processing From Object To Relational Schemas Based On A Parameterized Canonical Representation.* International Journal of Cooperative Information Systems, 1995.
3. Sujansky, W., *Methodological Review. Heterogeneous Database Integration in Biomedicine.* Journal of Biomedical Informatics, 2001. **34**: p. 285-298.
4. Dampney, C.N.G., G. Pegler, and M. Johnson. *Harmonising Health Information Models - A Critical Analysis of Current Practice*. in *Ninth National Health Informatics Conference*. 2001. Canberra ACT, Australia.
5. Lee, M.L. and R. Ramakrishnan, *Integration of Disparate Information Sources: A Short Survey.* ACM Multimedia, 1999.
6. Bakken, S., et al., *Toward Vocabulary Domain Specifications for Health Level 7-Coded Data Elements.* Journal of the American Medical Informatics Association, 2000. **7**(4): p. 333-342.
7. NLM, *Fact Sheet. UMLS ® Metathesaurus*. 2003, National Library of Medicine.
8. Snodgrass, R.T., *Developing Time-Oriented Database Applications in SQL*. The Morgan Kaufmann Series in Data Management Systems, ed. J. Gray. 2000, San Francisco: Morgan Kaufmann, Inc. 484.
9. Spyns, P., *Natural Language Processing in Medicine: An Overview.* Methods of Information in Medicine, 1996. **35**: p. 285-301.
10. Manning, C.D., *Foundations of Statistical Natural Language Processing*. 2000: Massachusetts Institute of Technology.
11. Baud, R. and P. Ruch, *The Future of Natural Language Processing for Biomedical Applications.* International Journal of Medical Informatics, 2002. **67**: p. 1-5.
12. Baud, R., A. Rassinoux, and J. Scherrer, *Natural Language Processing and Semantical Representation of Medical Texts.* Methods of Information in Medicine, 1992. **31**: p. 117-125.
13. Sycara, K., M. Klusch, and J. Lu. *Matchmaking Among Heterogeneous Agents on the Internet*. in *AI Spring Symposium on Artificial Agents on Cyberspace*. 1999.
14. Mori, A.R., et al. *Conceptual Schemata for Terminology: A Continuum from Headings to Values in Patient Records and Messages*. in *AMIA*. 2002.
15. Brodie, M.L. and M. Stonebraker, *Migrating Legacy Systems. Gateways, Interfaces & The Incremental Approach*. 1995, San Francisco: Morgan Kaufmann Publishers, Inc. 195.
16. Kroenke, D., *Database Processing. Fundamentals, Design & Implementation*. 1999, Upper Saddle River: Prentice Hall.
17. Huff, S. and J. Carter. *A Characterization of Terminology Models, Clinical Templates, Message Models, and Other Kinds of Clinical Information Models*. in *AMIA Symposium*. 2000.
18. Davis, R., H. Shrobe, and P. Szolovits, *What is a Knowledge Representation?* AI Magazine, 1993. **14**(1): p. 17-33.

19. Burgun, A., et al., *Issues in the Design of Medical Ontologies Used for Knowedge Sharing.* Journal of Medical Systems, 2001. **25**(2): p. 95-108.
20. Kayed, A. and R.M. Colomb, *Extracting Ontological Concepts for Tendering Conceptual Structures.* Data & Knowledge Engineering, 2002. **40**: p. 71-89.
21. Nodine, M., J. Fowler, and B. Perry. *Active Information Gathering in InfoSleuth*. in *International Symposium on Cooperative Database Systems for Advanced Application*. 1999.
22. Ramakrishnan, R. and A. Silberschatz, *Scalable Integration of Data Collection on the Web. Technical Report.* 1998, University of Wisconsin - Madison: Madison.
23. Sugumaran, V. and V.C. Storey, *Ontologies for Conceptual Modeling: Their Creation, Use, and Management.* Data & Knowledge Engineering, 2002. **42**: p. 251-271.
24. Fridman, N. and M.A. Musen, *PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment.* 2000.
25. McGuinness, D.L., et al., *The Chimaera Ontology Environment.* American Association for Artificial Intelligence, 2000.
26. McGuinness, D.L., *Conceptual Modeling for Distributed Ontology Environments.* Proceedings of the Eighth International Conference on Conceptual Structures Logical, Linguistic, and Computational Issues (ICCS 2000), 2000(August 14-18, 2000).
27. Mena, E., et al., *Observer: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies.* International Journal Distributed and Parallel Databases, 1998.
28. Mitra, P., G. Wiederhold , and M. Kersten. *A Graph-Oriented Model for Articulation of Ontology Interdependencies*. in *Proceedings of Conference on Extending Database Technology*. 2000. Konstanz, Germany.
29. Spackman, K.A., K.E. Campbell, and R.A. Côté, *SNOMED RT: A Reference Terminology for Health Care.* Journal of the American Medical Informatics Association, 1997. **4(Symposium Supplement)**: p. 640-644.
30. ISO, *2000 Terminology Work - Vocabulary - Part 1: Theory and Application (Final Draft International Standard)*. 2000, International Organization for Standardization.
31. Elkin, P.L., et al., *Guideline and Quality Indicators for Development, Purchase and Use of Controlled Health Vocabularies.* International Journal of Medical Informatics, 2002. **68**(1-3): p. 175-186.
32. Neces, R., et al., *Enabling Technology for Knowledge Sharing.* AI Magazine, 1991. **12**(3).
33. Campbell, K.E., D.E. Oliver, and E.H. Shortliffe, *The Unified Medical Language System: Toward a Collaborative Approach for solving terminologic problems.* Journal of the American Medical Informatics Association, 1998. **5**(1): p. 12-16.
34. Cimino, J., *Desiderata for Controlled Medical Vocabularies in the Twenty-First Century.* Methods of Information in Medicine, 1998.

Creating Homogeneous Data from Heterogeneous Representations: A Process for Heterogeneous Database Integration*

Rachel L. Richesson, PhD, MPH[a,*], James P. Turley, RN, PhD[a], Kathy A. Johnson-Throop, PhD[a], Christoph Eick[b], PhD, Mark S. Tuttle, FACMI[c]
[a]University of Texas Health Science Center at Houston, School of Health Information Sciences, [b]University of Houston, Department of Computer Science, [c]Apelon, Inc.

*[SUBMITTED TO DATA AND KNOWLEDGE ENGINEERING, AUGUST 2003]

Abstract:

The goal of heterogeneous database integration is to achieve a homogeneous representation for the comparability of the underlying data. A generalized process for creating a homogeneous representation while preserving local data granularity and intended meaning was developed within the context of a heterogeneous database integration problem in the health care domain. This process includes the creation of a global schema, supporting reference terminologies, and the representation of important characteristics related to the quality and precision of local term–reference terminology concept mappings. This process addresses common problems arising from heterogeneous databases and can be generalized to other domains.

Keywords:
Heterogeneous database integration; data manipulation; knowledge representation; data quality; database construction

***Corresponding Author:**
Rachel Richesson, PhD, MPH
University of Texas Health Science Center at Houston
School of Health Information Sciences
7000 Fannin, Suite 600
Houston, TX 77030 USA
713-500-3456, 713-500-3915 (FAX)
Rachel.L.Richesson@uth.tmc.edu

1.) Introduction

Any sensible use of data from heterogeneous databases requires a uniform or homogeneous representation for comparability. Integrating heterogeneous databases into a homogeneous representation, while preserving the intended semantic meaning and data granularity from native representations, presents both a conceptual and practical challenge. The focus of this research is on the development of a generalizable process for heterogeneous database integration that achieves comparable data with a homogeneous representation. The process was developed by systematically addressing common classes of representational heterogeneity and resultant data conflicts [4] exhibited by heterogeneous databases using a data integration project in the health care domain.

2.) Related research

This section describes the current research and outstanding issues involved in the integration of heterogeneous databases.

2.1 Heterogeneous databases

Heterogeneous databases can be defined as separate autonomous databases, independently created for unique purposes, with substantial differences in both abstract data models and database schema.[1] The importance of integrating heterogeneous databases is illustrated by the great number of research review articles on the subject,

across many disciplines.[1, 8, 40-42] The challenge for creating integrated data with a uniform, or homogeneous, representation is in identifying and resolving all of the heterogeneities, or differences, that exist between the source databases. Heterogeneity from multiple autonomous databases arises from representational differences that manifest in a variety of data conflicts, many semantic in nature. These representational differences include naming and formatting differences in attribute names, structural differences in table and attribute decompositions, and semantic differences in the definitions of data attributes and underlying data content.[4] The goal for integrating heterogeneous databases is to resolve these representational differences to a uniform representation; ideal solutions preserve intended meaning and granularity from local sources.[4]

2.2 Semantic data conflicts

Representational heterogeneities can result in differences in data semantics that can impact the quality of the data for secondary use. Broadly, semantic differences or data conflicts (also called semantic heterogeneities) occur when there is a disagreement about the meaning, interpretation, or intended use of the same or related data, and arise from different definitions of data attributes, differences in coding precision of the data content across multiple databases [1], or context [9]. Broadly, semantic data conflicts arise when the data in different systems is subject to different interpretations, even when data types, labels, and general schemas are identical.[10] For example, three separate restaurant-review databases might each contain an attribute called “meal cost”, yet the

meaning of the construct may differ in each source. One database might use meal cost to mean the menu cost, one might use meal cost to mean the cost of the meal including tip, and another might use it to mean the cost of the meal including tip and tax. Therefore, the underlying data in these attributes would not be comparable, despite having similar attribute names and data definitions, because the intended semantics differ for each. Data from different source representations cannot be validly compared or aggregated without assurance that the semantic intent of each data value is understood. While such semantic data conflicts are often difficult to precisely define, identify, and classify [1], there is common consensus that their resolution is the most problematic aspect of heterogeneous database integration efforts.[8] [10]

2.3 Heterogeneities caused by different concept and measurement systems

A common data integrity challenge for heterogeneous database integration efforts is the assurance that the concept and measurement systems encoding the data are *comparable* across attributes that need to be aggregated. Comparability is a broader notion than equivalence, and implies the need for a common representation, or standard, to make judgments of relationships between different values (e.g., equivalent to, greater than, less than; broader than, narrower than, etc.). Classic examples of measurement system differences include length in feet vs. inches, or weight in pounds vs. kilograms. Since the conversions for such different ratio systems are well-known, these disparities

reduce to a common problem of scaling, resolved by simple re-coding to a standard measurement system.

Commonly, however, data content is encoded in knowledge representation systems (e.g., terminologies, coding schemes) that represent concepts as the "units" of measure or membership, and heterogeneities between different representations of data attribute content can be problematic to resolve. While the need for a standard measurement system or a reference model is required to resolve differences in classic measurement systems, conversions between different *concept systems* requires an understanding of the conceptual "units" of each system. For example, what is the relationship between "coughing/wheezing" in one concept system and "breathing problems" in another? Or, what is the relationship between "nasal congestion" in one concept system and "runny nose" in another? The transformation of different concept systems to a standard ultimately requires an understanding of the intended meaning of each local data value, and its relationship to selected standard concepts. Such concept systems do not evaluate constructs on a continuous or numerical scale, as do classic measurement systems, but evaluate the membership of a given instance in a concept or class of concepts. Failure to resolve such data values at the conceptual level can result in potentially serious precision data conflicts and confounding of meaning.[8] The existence of many concept systems for health care knowledge is frequently referred to as the "vocabulary problem" in the medical informatics literature, is viewed inherently as a knowledge representation problem, and is a major research focus for the field.[11]

One solution for achieving a homogenous representation of heterogeneous data content is to map concepts from different concept systems to a standard or referent concept system, called a *reference terminology*. A reference terminology is a specified set of concepts and relationships that provides a common reference point for the comparison and subsequent integration of heterogeneous data. [34] The mapping, or transformation, of the concepts underlying heterogeneous local terms to a standard conceptual framework creates comparability of data content, and facilitates the integration, storage, and retrieval of data from multiple sources. Unlike the use of implicit reference models or pair-wise comparisons, the use of a singular reference terminology can be explicit, open to evaluation, and can integrate more than 2 concept systems.

Reference terminologies can be created from one or all of the component terminologies, borrowed, modified, or created a new. The fitness of a reference terminology is entirely dependent upon the purpose [36], and it has been said that for any purpose, no perfect terminology exists.[37] The ideal reference terminology should have the concept coverage and (i.e., specificity or detail) to meet the intended needs and to capture local data.[39] Since reference terminologies are indeed knowledge representation systems, the literature on conceptual modeling and ontology development provides good resources for design strategies. All of these sources emphasize the highly iterative development process, the reliance on domain experts, literature, and data instances for development and refinement, and the importance of the purpose for which these conceptualizations are created to support. While there are few standards for the design of reference terminologies per se, there are established guidelines for the

evaluation of terminologies in healthcare that can be generalizable in part to other domains.[36, 39]

2.4 Integration of heterogeneous databases

The transformation of heterogeneous databases into a homogeneous representation facilitates the querying or extraction of data from the component databases. Two broad approaches are used to query data from heterogeneous databases: data-translation (i.e., data integration and summarization) and query-modification. The first strategy extracts the desired data from each source using the query language specific to each database, and then translates the data from each source to the standard or uniform representation. This 'data-translation' strategy is typically used in data warehouse or clinical data repository projects. [8] The second approach involves translating a desired query into equivalent functional queries for each local data source to extract data from each source. Such 'query-translation' strategies involve information mediators or "wrappers" for each local system that describe what the databases can provide in terms of the local abstract data model and database schema, and what types of queries they can answer in terms of the native query language.[18] In both strategies, users are oblivious to any representational differences across component databases. Similarly, users are usually unaware of the precision of the transformed homogeneous data relative to its original representation. To ensure the integrity and to understand the limitations of the transformed data, both the data-translation and query-translation approaches ultimately

require the same thorough examination of the intended meaning of each heterogeneous database structure and the underlying concept and measurement systems.

Data from heterogeneous databases is comparable if the semantic intent of the underlying data can be transformed to a homogeneous or standard representation. This standard representation is considered “global” relative to the local component data sources, and consists of a standard database schema representation as well as standard representations for the supporting concept and measurement systems encoding the data content. Once standard representations are selected, the data integration process entails ‘mappings’, or defined relationships, between both local data structures and underlying content, to the global knowledge representations. These standard representations can support both data-translation and query-translation database integration approaches.

3.) Creating homogeneous data from heterogeneous databases

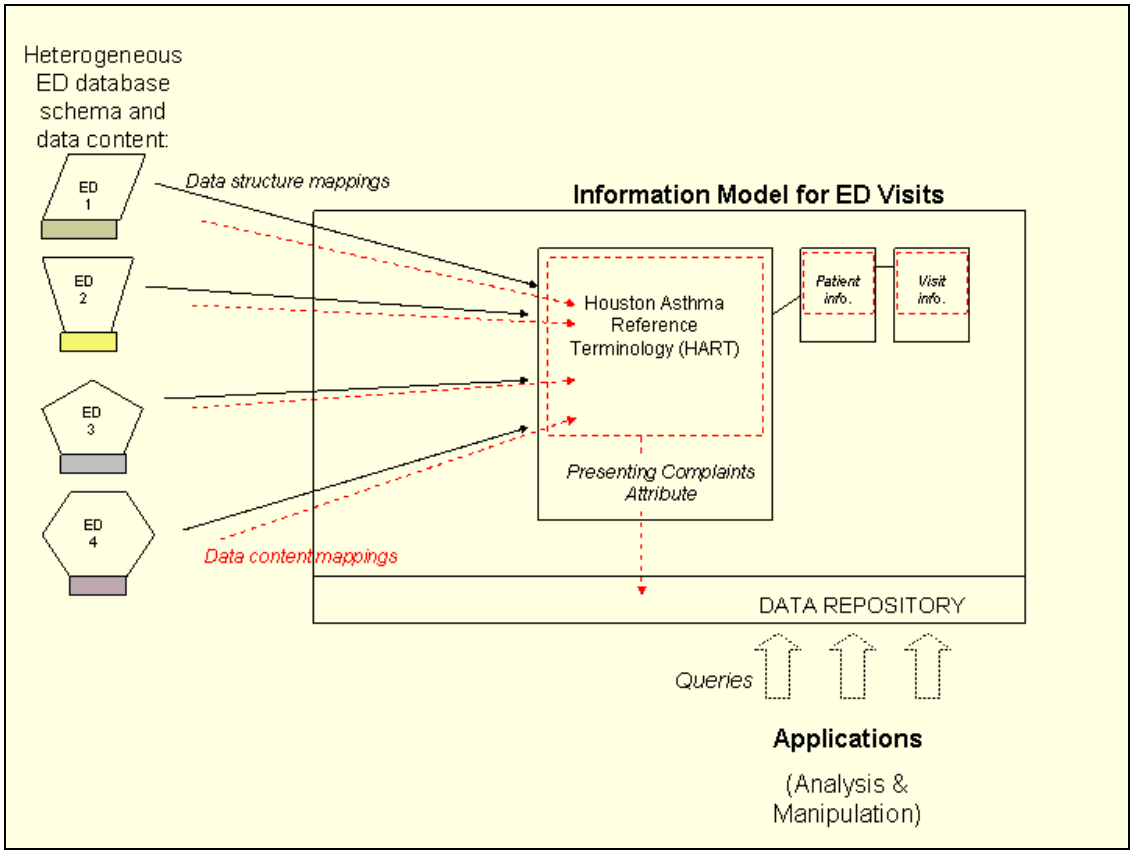
Successful data integration solutions preserve data granularity and intended meaning of local data values while transforming them into a homogeneous representation for comparability and compilation. A database integration effort in the health care domain systematically identified and addressed representational heterogeneities across databases [4] to develop a generalizable process for creating a homogeneous data representation in a central data repository. The goals of this section are to describe specific requisites for valid data integration from both the literature and this data

integration experience. These requisites include the use of conceptually-based reference standards (i.e., global schema and reference terminologies), characteristics of the mapping relationships between native and standard data representations, and the role of context in preserving data integrity. Section 4 assimilates these issues into a generalizable process for achieving homogeneous data from heterogeneous database representations.

3.1 General strategy for creating uniform representation

This project transformed presenting complaint (i.e., patient-reported reason for visit) data from 4 different hospital emergency departments (EDs), disparate in both database schema and underlying content, to a standard representation. (Figure 1) This standard representation includes a global database schema and a supporting reference terminology to homogeneously represent heterogeneous database schema and data content. Heterogeneous presenting complaint data content was made homogeneous by mapping disparate presenting complaint data values to a reference terminology for presenting complaints, called the Houston Asthma Reference Terminology (HART), that was created especially for this project.[2, 6] The HART provides a uniform representation for presenting complaint data from the 4 heterogeneous sources, and represents a concept system that encodes one attribute (presenting complaints) of a global database schema for ED visits. This global database schema gives a uniform context to the homogeneous presenting complaint data, and provides the structure for a data repository, which can be queried to support a variety of public health and research applications.

Figure 1. HART Project Overview



Prior to the start of this data integration, domain experts were queried to determine the purpose and scope for the database integration, and to share insight into the global conceptualizations of the domain. The database schema and data content from the local or component databases were systematically examined, first individually to understand content and structure, and then relative to the other component databases to understand content and representational similarities and heterogeneities between them. An understanding of the data content and structure from the component data sources, as well as the intended purpose of the final data, drove the development of reference

standards (global schema and reference terminology) that provided the structure and representation for the final repository data.

Key activities for obtaining homogeneous data while preserving semantic intent and data granularity identified through the literature and this data integration include: the development of conceptually-based reference standards (i.e., global schema and reference terminologies), creating mapping relations or assertions between local data structures and referent standards, identifying relevant mapping characteristics, and identifying contextual elements that influence local data semantics. These key activities are each discussed in this section, and are then assimilated into the generalizable process for achieving homogeneous data from heterogeneous database representations that is outlined in Section 4.

3.2 Conceptually-based reference standards

Only when data share a common representation can they be compared. Integration of heterogeneous databases requires one or more standard representations to resolve disparities in both database schema and the underlying concept systems encoding the data. In order to preserve the intended meaning of each local data structure, the standard representations should conceptually-based, meaning that the units of information are concepts or groups of concepts.[4] In the resolution of heterogeneous database schema, the standard representation is called a *global database schema*.[1, 10] In the resolution of knowledge representation systems for data attribute content (e.g., concept systems), the

standard representation is called a *reference terminology*. Together, global database schema and reference terminologies constitute the homogeneous representation structure that facilitates the secondary use of existing data. Mappings between local concepts or data structures and the reference terminology and global schema constructs transform the heterogeneously represented source data to this homogeneous representation.

3.2.1 Global Schemas

Most current general processes for heterogeneous database integration involve some type of transformation of local database schema, often to a master or global schema. This global schema defines all of the important data structures and relationships required at the aggregate level, and as such, it forms the limits of what the new data attributes and relations can express. The global schema guides the query of the repository (in the data-translation approach) or defines the translations, or mapping, of component heterogeneous query models to the referent in the query-modification approach. The global schema can be thought of as the schema of a final integrated database, or the "ideal" schema in terms of the final purpose.

A global schema can be created from two different approaches: a bottom-up *schema integration* (literally combining schema from existing heterogeneous databases), or a top-down *schema creation* (driven by a broader conceptual organization or purpose-driven view). While both approaches have their merits, the construction of a global schema is often highly iterative and therefore involves some element of both approaches. Methodologies for the development of ontologies provide good resource for the design of

global schema.[24, 25, 28-31] The use of ontologies, defined as global conceptualizations of domain knowledge, can facilitate the development of global database schema. [18, 24-27]

In this project, a global schema for ED visits was developed from domain experts' conceptualizations of the ED visit process. The basic model of ED care was confirmed by asking several experts about the process from both the patient's point of view and from a data collection perspective. The local data schema were examined, attribute by attribute, from each source to see if the local data attributes matched the conceptualization of ED visits given by experts. There was consensus on the nature of ED visits, which was supported by the structures of the local database schemas, and so the definition of the ED global schema developed very quickly. The conceptualization for ED visits that supported the development of this global schema is that patients (who have more or less permanent characteristics) present with complaints, have an acuity/severity, receive a diagnosis, and leave the ED with a final disposition status. The resulting global schema (relational model) divides the information from each ED visit into patient information (e.g., demographics, identifier) and visit-specific information (e.g., date/time of visit, hospital of visit, visit identifier). Each visit is related to one or more presenting complaint concepts, and one diagnosis, acuity, and disposition value. This global database schema for ED visits "normalizes" presenting complaint data from the local data attributes, creating a common data attribute for all presenting complaint data instances.

The mapping, or transformation, of local structures to the data attributes of the global schema was used to overcome naming, format, structural, and content differences across the native databases in this data integration. For example, local data structures (at the table, attribute or instance level) such as “chief complaint”, “presenting complaint”, and “PresCompl” were mapped to the master “Presenting Complaints” attribute in the global schema, yielding semantically-like data in a homogeneous representation. The global schema, as any other database schema, included attribute definitions and formats, and relationships between attributes, which guided the mapping of local data structures to global data structures, addressing many representational disparities. The global schema also helped to identify semantic data conflicts, by guiding the systematic investigation of operational data definitions across local sources. Domain experts and local database users were critical in this process to identify the intended and desired meaning of data structures and definitions in the local and global schema. The global schema provided a blueprint of what data definitions and relationships to investigate. Similarly, each data attribute of the global schema was examined and considered for the use of a reference terminology to represent content from any underlying heterogeneous concept systems.

The global schema guides the storage and retrieval of the transformed and integrated data. For this project the global schema was represented as attributes (tables) and relations in a relational (MS Access) database. The global schema can also be represented as the database structure in other applications using different database paradigms. The mapping between local database schema and global schema require programming activity, as data structures are being renamed and/or transformed, and data

relationships are being changed or created in the global schema. The programming language used in this demonstration was Visual Basic implemented within the Access database environment. However, any programming language with the capability to transform data structures (e.g., Python, C++, Java) would function. Specialized append queries could also be used to query structures from data tables and populate the attributes of other "master" tables.

3.2.2 Reference terminologies

The development of a global schema for ED visits guided the resolution of many representational and semantic data conflicts, but an additional knowledge representation standard, or reference terminology, was required to resolve differences in concept systems encoding the underlying presenting complaint data.

The development of the standard HART reference terminology to represent presenting complaint data from heterogeneous ED databases was highly iterative and required 6 person months of development time, including numerous observations and interviews with data administrators and domain experts. The development is discussed in detail elsewhere [2] but could be summarized as an iterative, conceptually-based data-driven process consisting of multiple iterations of the following key steps: specify purpose, list relevant concepts, describe relevant concepts, identify important inter-concept relationships, organize relevant concepts, test inter-concept relationships, and choose appropriate representation.

The lack of one-to-one correspondence between concepts represented by the local concept systems complicated the development of the HART, and forced choices to be made about the content and structure of the reference terminology.[2] Relationships between local terms such as “difficulty breathing” and “respiratory problems” had to be understood and expressed in the final reference terminology. The presence of coded terms representing multiple or “lumped” concepts influenced the reference terminology structure. For example, coded terms such as “cough/secretions” had to be dissected into distinct concepts of “cough” and “secretions”, and choices had to be made if such lumped terms should drive a similar category in the final HART reference terminology, which would result in some loss of granularity for concept systems that singularly represented the more granular concepts (e.g., cough). Local terms representing implied concepts also influenced the final HART structure and the associated local term-HART concept mappings. For example, the semantic intent of terms such as “Flu-like symptoms” and “cold” had to be explored, and distinct, yet implied, concepts such as fever, vomiting, and nasal congestion teased out.

The physical format of the local term-Reference Terminology concept mappings can vary by actual implementation environment (e.g., relational table of equivalencies, different types of programming syntax), but the mappings all represent assertions for valid data transformations. In this demonstration, relationships were specified between local concepts and HART concepts as part of the schema design of the final Access data repository. Regardless of the tools and languages used to transform local terms into a common concept terminology, the mappings are ultimately created by domain experts.

Because of the important influence of data collection context on semantic meaning, it is likely that the mapping experts will be different for each source database.

3.3 Characteristics of mapping

The heterogeneous nature of the local data instances demonstrated variability in the local term-HART concept mapping relationships, and characterizations of this variability emerged in this data integration effort. Often the local term-HART concept mapping variability entailed a loss of granularity or intended meaning from some source representations. For example, if local terms “dry cough” and “cough” are both mapped to the concept “cough” in a reference terminology, the transformed data would appear similar (i.e., 2 instances of “cough”) but any distinctions between the local terms would be lost. The capture of the different relationships between term-concept mappings for both local terms, however, can allow some data granularity and potential meaning to be retained. Noting that the local term “dry cough” is more granular or specific than the reference concept “cough” and that the local term “cough” represents the same concept as reference concept “cough” can facilitate understanding of the relationship that exists between the two local terms. Most database integration efforts transform data to a common representation making the user unaware of any disparities in the native representations. Processes that address these *precision* differences in local term-referent concept mappings can enhance the end users’ understanding, querying, and use of final data. Making these mapping characteristics explicit has implications both for the

evaluation and maintenance of the reference terminology and for utility of the transformed data in computational or statistical analyses.

A representation of precision for this project was adapted from that used by the Unified Medical Language System, a metathesaurus of biomedical concept systems developed by the National Library of Medicine.[12] The final representation for the precision of each local term-concept mapping in this project includes: exact term and concept, lexical variation (same concept), synonym (i.e., same concept), broader than, narrower than, related concept. In the implementation of global schema for ED visits and supporting HART created here, each local term-concept mapping is associated with one attribute describing the precision of the match.

Quality can be defined as the truthfulness of the local term-HART concept mappings in preserving semantic intent. In this sense, precision is a measure of quality. There are other measures of quality, and the conceptual and operational definitions of these constructs will vary by application and domain. One additional measure of the quality of the mapping in this project is a representation of who created or validated a given mapping assertion. Ultimately, a 3rd party creates the term-concept mappings that transform each local term to a standard reference terminology concept in heterogeneous data integration efforts. The fact that a physician, or more specifically, a pulmonologist, reviewed the term-concept mappings was an important quality attribute for this project. Depending upon the application or use case, it might be of greater importance to know that the data entry clerk who actually coded a term validated its semantic intent. The best-

suited mapping experts will differ for the problem or information-type being examined. Explicitly representing this basic measure of quality of each term-concept mapping assertion can add power to the compiled data, facilitating the potential quantification of certainty about the match.

The final representation for quality developed for this project captures who asserted the local term-HART concept mapping (medical expert from ED where instance originated, medical expert from other ED, nurse coder from ED where instance originated, nurse coder from other ED, or health informatics developer). In the final repository (relational) database schema created for this project, each local term-concept mapping is associated with one attribute describing the quality of the mapping assertion.

3.4 Context

Consideration of context surrounding local data instances can help determine their intended meaning. In computing terminology, context is defined as that which surrounds, and gives meaning to, something else.[43] Explicit representations of context on multiple levels (e.g., data instance, database schema, data collection process, data collection quality, and domain) facilitate the development of uniform conceptual knowledge representations and their associated mappings.[44] The context of the data instance guided the mapping of local free-text terms in this data integration project. For example, the term “no cough” did not map to the HART referent concept “cough” because of the “no” characters surrounding the term of interest. In addition, the context of local database schema was used to distinguish semantic intent of the same term in different source

databases. For example, the coded term “Cough/Secretions” likely means either cough *or* secretions *or* both, whereas the free-text unstructured instance “cough, secretions” likely means both. In addition to data types, the context of local database schema was used to distinguish between transformed patient records with one presenting complaint concept due to constraints in the local database schema (i.e., the schema only allows one presenting complaint to be entered) versus the reality of the clinical situation (i.e., the patient truly had only one presenting complaint). To accomplish this, local data schemas were related to each data instance in the global schema for ED visits, which allowed traceability to native data formats and system constraints.

The context of local data collection settings and procedures is particularly important in identifying the semantic intent of local data values. Data collection context refers to organizational, setting, and process features that influence data collection and impact the semantic meaning of the data. Dampney et al. (2001) note that implicit or even obvious information at the database level is often not represented when taking data to an aggregated level.[9] Database schema typically do not explicitly code data that are implied or unnecessary for the database’s intended purpose (e.g., most physicians at a children’s hospital are pediatricians). For example, the general model of emergency care is that patients arrive, state one or more complaints, are triaged by a nurse, and then are diagnosed and treated by a physician. Although usually not represented in the database, it is known in the domain that a “presenting complaint” is reported by the patient. Therefore, a *presenting complaint* of asthma (reported by patient) carries a different meaning than a *diagnosis* of asthma (reported by a physician). Implicit information

structures typically not represented in health care databases include context of role, organization, purpose for which data are collected, and underlying concept systems.[9]

Identifying differences in data from context due to implied concepts or data collection procedures required domain experts and persons familiar with the context of data collection. Since the context differs for each data source, information to resolve disparities must be obtained from representatives familiar with both the conceptual database design and the routine coding procedures for each data source.[44, 45] The gathering of such information for this project required research techniques (structured interview and observation) routinely used in other attributes, such as psychology, sociology, and education. In addition, domain experts familiar with the broader domain knowledge and process models identified disparities in organizational and data collection context within the broader framework and needs of this data integration effort. It is likely that in other projects, domain experts will play a central role in identifying implied data attributes and/or data content.

4.) A generalized process

Based upon the ED presenting complaint data integration effort described above, along with the identification of key features of successful database integration solutions, a generalized process was developed for the integration of heterogeneous databases to a uniform representation. This generalized process accommodates the needs for the representation of quality and precision attributes that can preserve data granularity and

semantic meaning. As discussed in the previous section, this process is highly iterative and not as simplistically sequential as presented below. This generalized process and evaluation criteria target key sources of representational and semantic heterogeneity that challenge efforts to create a homogeneous representation in any domain. Despite the range of project requirements and variety in local database source representations in other potential database integration projects, successful efforts for integrating data from heterogeneous database and concept system representations into a homogeneous representation should include the following broad steps:

***1.) Define purpose, information needs, and process needs*.** This step is critical and should guide all choices to be made in the design of homogeneous representation standards and use. The purpose and information needs for the combined data dictate the level of detail and organizational structure required for the conceptual referents. Information needs are best represented by creating typical "use cases" that illustrate type, detail, and applications that the aggregated homogeneous data needs to support. Process needs (e.g., access, timing, data availability) drive the logistical procedures of the heterogeneous database integration. The purpose should be mutually defined and endorsed by a representative sample of potential application users, database integrators, and domain experts.

The model of database federation (i.e., the general approach for accessing and integrating data) is based upon the needs defined above, and determines the practical implementation of the database integration effort. The access permissions and

anticipated needs for updated or current data determine whether a query-modification or data-translation approach should be taken. In general, needs for current and frequently updated data are best satisfied with a query-modification approach, whereas data-translation is suitable for periodic data needs. Also, data sources with highly disparate concept systems requiring one or more reference terminologies will need a data-translation step. Specific guidance for selecting from different models of database federation can be found elsewhere.[1]

2.) Examine data structures, concept and measurement systems, and data collection context from local data sources. Each local data source must be explored to determine the semantic content. This examination can be *bottom-up,* meaning each local database is examined attribute by attribute, or *top-down,* meaning relevant constructs are identified from a conceptual model and the corresponding data structures or attributes are sought in each of the local database schema. Regardless of the approach, all relevant data attributes should be reviewed and synonymy in attribute names and definitions noted. In addition, the operationalized data definitions for each data attribute should be identified. The concept and measurement systems for each data attribute should be explored. This preliminary, almost qualitative, analysis of source database schema and underlying concept and measurement systems should identify each data attribute of interest in the final project and attempt to define initial equivalency relationships across databases. The level of disparities observed in structure, naming, format, data definitions, and concept or measurement systems encoding each data structure will dictate the best approach for defining the global

schema. Accordingly, the activities associated with this step are a prerequisite for step 3, defining the global schema.

3.) Define global schema. Depending upon the levels of disparities between the local databases and the overall project purposes, a schema integration approach (data-driven) or a top-down schema creation approach could be used. Regardless of the method, the global schema should define the constructs and relationships needed for the application, at both the level of granularity needed and with the terminology (attribute labels) familiar to the domain. The potential disparities in data attribute definitions across source databases and context of data collection that impact these final operational definitions, as well as important quality attributes identified by domain experts, should be represented in the global schema. Available domain ontologies (including conceptualizations of process and work-flow) should be searched for relevance and used as a resource to guide the development and/or refinement of the global schema. Ideal global schema should maintain relationships to the local data schema that allow the traceability of the native data context.

4.) Define reference terminologies and measurement systems. All concept and measurement systems supporting the final data attributes must have a standard representation. Each data structure in the global schema should be considered for a reference terminology that reflects expression and context of local data and encompasses representation needs for real use-cases. The level of disparity in concept systems across local databases and the information needs for the aggregated data

(both organization and granularity of content) determine whether these reference terminologies should be borrowed from other sources, or created via top-down or bottom-up approach. One reference terminology might capture the concepts represented in multiple data attributes of the global schema. The use of reference terminologies to achieve comparability from heterogeneous data can be termed *content integration,* and a generic process for this is described in [2].

5.) Map data structures expressed in local data sources to the closest construct in the global schema. This step generally can be thought of as reconciling local database attributes to corresponding attributes in the global schema, but can also involve moving data from instance level to attribute level to table level. Regardless, the meaning or concept class represented by each structure should be the focus of this activity. To preserve semantic intent from each local source, the focus goes beyond the data definitions of each database to include interviewing designers and users of each local database. Questions to be asked should include: *What is the meaning of this attribute? How is the content or value selected? Do all users agree? Does the context of the data collection influence the meaning of the data values? If so, how?* This examination of constructs at the local level might drive changes in the global schema. The ultimate information needs and purpose should guide the mapping of relevant local data structures to the appropriate structure in the global schema.

6.) Map relevant concepts that are implied but not explicit in local data models to the global schema. The global schema should identify constructs or structures to

compare to the local schema. Where constructs are missing, but are implied or can be derived from local sources, they must be imputed into the global schema. For example, if the construct of who recorded a particular local data attribute is important to the quality of the final data semantics, this concept should be included in the global schema and the values imputed appropriately by local users and domain experts. The project logistics determine how the imputation process should best occur. This can be achieved by a "blanket" imputation (e.g., all values for 'reported by' are the same for a given source) or by selective value-based imputation (e.g., presenting complaint in hospital A is a 'diagnosis' structure if it contains the term 'asthma'.) A representation for missing concepts that cannot be implied or derived should be included in the global schema. The analysis of disparities in data collection context, and review by domain experts and end-users, facilitates the identification, representation, and mapping of implied concepts from local databases to the global schema.

7.) Map terms (and the concepts they represent) expressed in local data sources to the closest concept in the appropriate reference terminology. Again, the focus of this mapping needs to be on the underlying concept or intended concept expressed as a term in local databases. Semantic intent is determined by observing representative coders for each local data source, as well as questioning coders, local database developers, and domain experts. The exploration of the context of data collection, as well as the roles and training and objectives of local data entry persons at each level can assist in understanding semantic intent.

8.) Characterize the quality of mapping of local data instances to concepts in reference terminology. The identification of potential context items that impact the quality of match is facilitated by domain experts and designers of local data sources, and also by a domain ontology, if available. An appropriate representation for the certainty of mapping of data structures to the global schema, as well as who asserted each mapping, should be developed. Any variability in the data definitions of constructs (e.g., one data attribute definition is operationalized differently than a corresponding construct in another local database schema) should be explicitly represented relative to each mapping relationship. Domain experts and end users should specify representations for quality that are useful and meaningful to the final applications.

9.) Characterize precision of mapping of local data instances to concepts in reference terminology. Similarly, the intended meaning of each term in each concept system should be mapped to the most appropriate and closest concept in the reference terminology, and characterizations of the local term-referent concept mappings should be represented. Variability in concept systems from the heterogeneous sources implies a loss of data granularity from some sources, and this loss of precision should be represented in the final global schema as appropriate to the needs of the final compiled data. Any variability in the data definitions of constructs (e.g., one data attribute definition is broader in scope or more inclusive than that of the corresponding construct in the reference terminology) should be explicitly represented relative to each mapping assertion.

The steps outlined above are highly iterative, and certain steps will entail repeating previous activities. The first step of identifying the needs and purposes for the data integration is one that will need to be re-visited at every step, and will be the lens through which the evaluation of the process and the resultant homogeneous data is ultimately determined.

While the process of validly integrating data from heterogeneous databases into a common representation is highly dependent upon the domain, purpose, and local contexts, the overall themes and goals and measures of success are similar. Data from heterogeneous databases cannot be validly aggregated or compared without a uniform or homogeneous representation. The ultimate goal for heterogeneous database integration is to create “comparable” data in an organized representation suitable for a given purpose. Given the final information needs, the most successful heterogeneous data integration strategy should retain as much granularity from as many sources as possible. The evaluation of success is addressed by examining the nature of the final homogeneous representation, and the relationships between local data structures and those of the final representation.

Some broad questions can be asked to evaluate the utility and success of the process:

1.) Have the purposes for the integration been clearly identified and specific use cases created?

2.) Does the global data schema include all of the constructs and concepts and relationships required to meet the use cases?

3.) Have the measurement systems for each data structure or construct in the global schema been transformed to a standard measurement system? Have the concept systems for each data structure or construct in the global schema been transformed to a reference terminology or standard concept system? Has each data structure or construct in the global schema been considered for the use of a reference terminology? Do the measurement systems and reference terminologies selected for each final data attribute have the granularity and detail needed to support the final use cases? Do the measurement systems and reference terminologies selected for each construct limit the loss of data granularity from each local data source? Is the loss of data granularity acceptable to the final purpose and use cases for the aggregated homogeneous data? Is there an explicit representation for the quality of the mapping of local data values to each reference terminology? Is this representation useful in explaining the variability of mappings and loss of granularity?

4.) Has each local data attribute mapping to the global schema been examined? Have the operational data definitions for each been examined? Have interviews been conducted with a representative sample of database developers and users for each local database? Was the context of data collection observed for disparities in the operationalization of data definitions? Is there an explicit representation for the quality of the mapping of local data structures to each structure in the global

schema? Is this representation useful in explaining the variability of mappings and loss of granularity?

5.) Have domain experts identified important elements of context that might impact quality of the final data? Is this context explicitly represented in the final global data schema? Can a concept or measurement system be identified to represent the content of relevant context attributes?

6.) Does the final global data schema provide relationships to native or local database schemas for each of the local databases? Is missing data represented differently for missing data instances versus missing data constructs?

In general, the evaluation of the process of assimilating heterogeneously represented data to a uniform or homogeneous representation is relative to satisfying the needs of the intended purposes of the integration project. The generalized process and evaluation criteria described above ensure a systematic approach for the examination of all local data structures, the use of conceptual referent standards, and the evolving relationships between the two. The intended purposes and final information needs drive each iteration of this process as well as define its completion and success.

5.) Conclusions and future work

Creating homogeneous data from heterogeneous database representations is not a trivial task, but is necessary for any meaningful secondary use of the data. Just as there is a wide variety in data integration projects by domain, scope, content, and purpose, the

resources required for these endeavors vary. Successful data integration projects will require systematic investigation of the intended meaning of local data structures and content, and will apply research techniques such as structured interview and observation from other disciplines.

The contribution of this work is a generalized process for creating homogeneous or comparable data from heterogeneous data representations. Once a homogeneous data representation is achieved, the data from heterogeneous databases can be compiled, shared, manipulated, and leveraged to address a multitude of business requirements. This process recognizes that the differences inherent in native data representation and data collection contexts imply some loss of meaning or precision when being transformed to standard homogeneous representations. Key steps of this process attempt to minimize the loss of data semantics and granularity, potentially allowing better "quality" data in the final representation. This generalized process should be valuable to a number of database integration efforts in a number of domains.

AKNOWLEDGEMENTS:
The authors wish to thank Dr. Charles Macias from the Department of Pediatric Emergency Medicine at the Baylor College of Medicine and Dr. Marianna Sockrider from the Department of Pediatric Pulmonology at the Baylor College of Medicine for their expertise, assistance and enthusiasm throughout this project.

This research was facilitated by the Robert Wood Johnson Foundation: Managing Pediatric Asthma: Emergency Department Demonstration Program; Pediatric Texas Emergency Department Asthma Surveillance (TEDAS).

Rachel Richesson is funded by National Library of Medicine Fellowship in Applied Informatics #1 F32 LM07188-01A1.

References

1. S. Bakken, K.E. Campbell, J.J. Cimino, H.S. M., and W.E. Hammond, *Toward Vocabulary Domain Specifications for Health Level 7-Coded Data Elements,* Journal of the American Medical Informatics Association **7** (4) (2000) 333-342.
2. A. Burgun, G. Botti, M. Fieschi, and P. Le Beux, *Issues in the Design of Medical Ontologies Used for Knowedge Sharing,* Journal of Medical Systems **25** (2) (2001) 95-108.
3. J. Cimino, *Desiderata for Controlled Medical Vocabularies in the Twenty-First Century,* Methods of Information in Medicine (1998).
4. C.N.G. Dampney, G. Pegler, and M. Johnson. *Harmonising Health Information Models - A Critical Analysis of Current Practice*. in *Ninth National Health Informatics Conference*. 2001. Canberra ACT, Australia.
5. P.L. Elkin, S.H. Brown, J.S. Carter, B.A. Bauer, D. Wahner-Roedler, L. Bergstrom, M. Pittelkow, and C. Rosse, *Guideline and Quality Indicators for Development, Purchase and Use of Controlled Health Vocabularies,* International Journal of Medical Informatics **68** (1-3) (2002) 175-186.
6. N. Fridman and M.A. Musen, *PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment,* (2000).
7. D. Howe, *The Free On-line Dictionary of Computing*. 1993-2001, Denis Howe.
8. R. Jakobovits, *Integrating Heterogeneous Autonomous Information Sources,* (1997).
9. A. Kayed and R.M. Colomb, *Extracting Ontological Concepts for Tendering Conceptual Structures,* Data & Knowledge Engineering **40** (2002) 71-89.
10. M.L. Lee and R. Ramakrishnan, *Integration of Disparate Information Sources: A Short Survey,* ACM Multimedia (1999).
11. D.L. McGuinness, *Conceptual Modeling for Distributed Ontology Environments,* Proceedings of the Eighth International Conference on Conceptual Structures Logical, Linguistic, and Computational Issues (ICCS 2000) (August 14-18, 2000) (2000).
12. D.L. McGuinness, R. Fikes, J. Rice, and S. Wilder, *The Chimaera Ontology Environment,* American Association for Artificial Intelligence (2000).
13. E. Mena, V. Kashyap, A. Sheth, and A. llarramendi, *OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies,* Conference on Cooperative Information Systems (1996).
14. R. Neces, R. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator, and W.R. Swartout, *Enabling Technology for Knowledge Sharing,* AI Magazine **12** (3) (1991).
15. NLM, *Fact Sheet. UMLS ® Metathesaurus*. 2003, National Library of Medicine.
16. M. Nodine, J. Fowler, and B. Perry. *Active Information Gathering in InfoSleuth*. in *International Symposium on Cooperative Database Systems for Advanced Application*. 1999.
17. R. Ramakrishnan and A. Silberschatz, *Scalable Integration of Data Collection on the Web. Technical Report.* 1998, University of Wisconsin - Madison: Madison.
18. M.P. Reddy, B.E. Prasad, and P.G. Reddy, *Query Processing in Heterogeneous Distributed Database Management Systems,* (1988).

19. R.L. Richesson, J.P. Turley, K.A. Johnson, M.S. Tuttle, and C. Eick, *Heterogeneous Database Integration: Resolving Representational and Semantic Heterogeneity to Achieve Homogeneous Aggregate Data,* manuscript in progress (2003).
20. R.L. Richesson, J.P. Turley, K.A. Johnson-Throop, C. Eick, M. Sockrider, C.G. Macias, and M.S. Tuttle, *Obtaining Comparable Presenting Complaint Data From Heterogeneous Emergency Department Databases,* Submitted to: Journal of the American Medical Informatics Association (2003).
21. R.L. Richesson, J.P. Turley, K.A. Johnson-Throop, C. Eick, and M.S. Tuttle, *Development and Use of a Reference Terminology to Maintain Data Granularity and Semantics in the Integration of Heterogeneous Data Content,* Submitted to: Data and Knowledge Engineering (2003).
22. R.L. Richesson, J.P. Turley, K.A. Johnson-Throop, C. Eick, and M.S. Tuttle, *Foundations for Heterogeneous Database Integration: A Framework to Identify Representational Heterogeneities,* Submitted to: CACM (2003).
23. A.P. Sheth and J.A. Larson, *Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases,* ACM Computing Surveys **22** (3) (1990) 183-236.
24. K.A. Spackman, K.E. Campbell, and R.A. Côté, *SNOMED RT: A Reference Terminology for Health Care*, Journal of the American Medical Informatics Association **4(Symposium Supplement)** (1997) 640-644.
25. V. Sugumaran and V.C. Storey, *Ontologies for Conceptual Modeling: Their Creation, Use, and Management*, Data & Knowledge Engineering **42** (2002) 251-271.
26. W. Sujansky, *Methodological Review. Heterogeneous Database Integration in Biomedicine*, Journal of Biomedical Informatics **34** (2001) 285-298.
27. K. Sycara, M. Klusch, and J. Lu. *Matchmaking Among Heterogeneous Agents on the Internet*. in *AI Spring Symposium on Artificial Agents on Cyberspace*. 1999.
28. J.P. Turley, R.L. Richesson, K.A. Johnson-Throop, C. Eick, and M.S. Tuttle. *The Role of Context in the Integration of Heterogeneous Health Care Databases*. *Submitted to: American Medical Informatics Annual Symposium*. 2004.

Development and Use of a Reference Terminology to Maintain Data Granularity and Semantics in the Integration of Heterogeneous Data Content*

Rachel L. Richesson, MS, MPH[1], James P. Turley, RN, PhD[1], Kathy A. Johnson-Throop, PhD[1], Christoph Eick[2], PhD, Mark S. Tuttle, FACMI[3]
[1]University of Texas Health Science Center at Houston, School of Health Information Sciences, [2]University of Houston, Department of Computer Science, [3]Apelon, Inc.

*[SUBMITTED TO DATA AND KNOWLEDGE ENGINEERING, AUGUST 2003]

Abstract:

To create comparable data from heterogeneous databases, a common representation for both database structure (i.e., database schema) and underlying content must be used. This common representation functions as a reference model or standard to which the component data structures and content are mapped or transformed. Global database schema are the reference models or standards that enable the integration of heterogeneous database schema, and reference terminologies are the reference models that enable the integration of heterogeneous data content from disparate knowledge representation or concept systems. A reference terminology provides a common representation, and dictates how data from heterogeneous representations can be compared, aggregated, or manipulated. The development of a reference terminology is neither trivial nor exact, and is heavily influenced by both the local data representations and the intended uses of the integrated data. Heterogeneous data often maps to the reference terminology with varying levels of precision and quality. The explicit representation of these variable mapping relationships enhances the use of a reference terminology and has implications for both the quality of the integrated data and the maintenance of the reference terminology. A process was developed to create a uniform representation for heterogeneous data content while preserving local data granularity and intended meaning. This process includes the development of a reference terminology and the representation of important characteristics related to the quality and precision of local term–reference terminology concept mappings. This process was developed within the context of a database integration problem in the domain of pediatric emergency medicine, and is generalized for other uses at the conclusion of this paper.

Keywords:

Heterogeneous database integration; content integration; knowledge representation; reference models; reference terminologies; measurement systems; concept systems; data quality

1.) Introduction

One prerequisite for data integration is that data from different sources be interpreted and integrated with respect to a common representation. Resolution of

heterogeneous data to a common representation involves resolving representational disparities in both database schema and the underlying measurement or concept systems (e.g., terminologies or coding schemes) encoding the data. Both require conceptually-based referents or standards to which the component data structures (i.e., tables, attributes, relations) or underlying data content from component data instances are mapped. *Global schema* are reference models that guide the integration of heterogeneous database schema into a uniform representation, and *reference terminologies* are reference models that guide the integration of heterogeneous data content into a uniform data representation. Reference terminologies are sets of concepts and relationships (also called knowledge representation frameworks or concept systems), that, when used as a standard representation, create "comparability", and therefore enable any subsequent aggregation, integration, communication, manipulation, and meaningful analyses.

The focus of this research is *content integration*, or the resolution of disparate concept systems that encode the data contained within heterogeneous database systems. Meaningful content integration solutions demand the examination of concepts underlying the terms or codes from each different concept system. The method is to transform concepts attached to local data instances to a single reference terminology, while preserving the granularity (i.e., detail) and intended semantics from local data instances. The contribution of this research is a process for the development of a reference terminology to uniformly represent data from heterogeneous sources, and explicit characterizations of the mapping relationships between native and final data representations. This process fits into a broader process for heterogeneous database integration presented elsewhere. [5] The development of this process in the field of pediatric emergency medicine is described, and a generalized process for integrating data from disparate representations is presented in the last section.

2.) Background

This section describes the outstanding issues involved in the integration of heterogeneous concept and measurement systems encoding the data from heterogeneous database representations. The use of standard knowledge representations (e.g., reference terminologies), to which heterogeneous data structures are mapped, is also described.

2.1 Integration of heterogeneous databases

The integration of heterogeneous databases requires the identification and resolution of many types of representational heterogeneities, or differences, between the source databases.[4] Many of these representational heterogeneities are due to differences in local database schema and are addressed by the use of a global, or referent, database schema. Many heterogeneous database integration projects have proposed methods for developing global database schema [7, 10, 33], and strategies for mapping heterogeneous database schema or data attributes to a referent global schema [1, 7]. However, little of the database integration research focuses on the resolution of disparate concept systems

or knowledge representations *within* data attributes, or on the measurement of how the component or native structures might relate (or "map") to the referent.

2.2 Representations for data content

The data contained in each database attribute is represented, either implicitly or explicitly, in a knowledge representation framework that can take several different forms. We broadly categorize these knowledge representation systems as one of two types: measurement systems and concept systems. Measurement systems are used to code data by evaluating *values* on a continuous or numerical scale, whereas concept systems encode data values or content by evaluating the *membership* of given data instance in a class, usually via the presence or absence of certain descriptive properties or characteristics. All concept systems include concepts and defined relationships between them. The simplest type of concept system is a coding system. A coding system is a combination of a set of concepts or rubrics (text string that describes a classing in a coding system or terminology), a set of code values, and a coding scheme that maps between the two. [35] Other types of concept systems increase in the complexity and formalized representation of semantic relationships, and include taxonomies, vocabularies, terminologies, and ontologies.

While the data encoded in a concept system is fundamentally different than the data represented in a measurement system, there is some overlap between the paradigms of measurement and concept systems. This is illustrated in nominal or ordinal measurement systems whose coding schemes either explicitly or implicitly represent concepts. The distinction between the four types of measurement scales (nominal, ordinal, interval and ratio) is based on the amount of information or the qualitative characteristics of the information carried by the data. Nominal and ordinal data can be represented by coding systems or terminologies, whereas ratio and interval data represent distinct points and ranges, respectively, on an absolute scale. Both coding systems and terminologies can be thought of as ordinal measurement systems, where the underlying units of measure are concepts, and are referred to here as types of concept systems.

2.3 Challenges for content integration

A common data integrity challenge for heterogeneous database integration efforts is the assurance that the concept and measurement systems are *comparable* across fields that need to be aggregated. Comparability is a broader notion than equivalence, and implies the need for a common representation to make judgments of relationships between the values (e.g., equivalent to, greater than, less than; broader than, narrower than, etc.). Classic examples of measurement system differences include length in feet vs. inches, or weight in pounds vs. kilograms. Since the conversions for such different ratio systems are well-known, these disparities reduce to a common problem of scaling, resolved by simple re-coding to a standard measurement system. Resolving disparate concept systems, including measurement systems involving nominal or ordinal data, however, are more challenging.

Whether heterogeneous database content is represented in a measurement or concept system, the process of content integration within heterogeneous databases requires one standard system as a reference model for each data attribute. In resolving heterogeneous measurement systems, the reference model is a standard measurement system; in resolving heterogeneous concept systems, the reference model can be either a standard concept system or a standard measurement system. The integration of different concept systems, can result in potentially serious precision data conflicts and confounding of meaning. [8] Resolving disparate concept systems requires examining the concepts underlying each local value, term, or coded representation, and transforming (or mapping) those to the appropriate concepts in a standard concept system. The transformation of concept systems to a standard system, either a concept or measurement system, therefore, is much more complex than the transformation of data from one measurement system to another. These (concept-concept) transformations are one focus of this paper.

Three different representations for measuring patient acuity (a measure of the seriousness of patient's condition) are shown in Figure 1. The granularity (i.e., level of precision and detail) of these scales differ: two of the coding systems (Emergency Departments A and C) represent acuity on a 3-value ordinal scale, while another (Emergency Department B) represents this same construct on a 4-value ordinal scale. Each scale represents some mix of patient and organizational characteristics that collectively classify the severity of the patient and their triage priority. Even without an understanding what concepts the specific code values represent, it is apparent that the use of any one of these concept systems as a standard coding will result in either the loss of data granularity from some coding systems, or the need to impute concepts from others.

Figure 1. Alternative Concept Systems for Patient Acuity Information

EMEDGENCY ROOM A	EMEDGENCY ROOM B	EMEDGENCY ROOM C
ASAP	Red	Team
Urgent	Blue	Check
Stable	Yellow	Shock
	Green	

Semantic data conflicts arise in heterogeneous databases when the units of measurement are not comparable for similar structures across databases, and this lack of comparability is often difficult to detect. It might seem logical to assimilate the different 3-value scales for Emergency Departments A and C, but if the code values "ASAP", "Urgent" and "Stable" do not have the same underlying concepts as "Team", "Check", and "Shock", this would lead to a semantic mis-match, or confounding of meaning, in the compiled data. Strategies to quantify these codes to a common representation, discussed further in the next paragraph, attempt to capture, at least implicitly, the semantic intent

underlying each code. The success of this capture of the true semantics of each term ultimately determines the validity of each numerical transformation.

Possible strategies to assimilate heterogeneous concept systems, such as those shown in Figure 1, include quantification and translation. In the quantification approach, the concept systems are transformed to a standard measurement system. For example, each local code would be assigned a numerical value or range based upon concepts or properties denoted or connoted by each data instance. The numerical codes would then serve as the reference standard or uniform representation for the content integration. Several types of measurement systems could be used as the standard in the quantification approach, including absolute and interval scales, and ranked or ordered scales. Ultimately, the selection and transformation of concept systems to a standard measurement system all require domain experts' guidance based upon their understanding of the concepts and clinical situations represented by each local code. In the quantification approach, however, the transformations of data from local codes to standard numerical values or ranges, do not explicitly represent the underlying concepts that are synthesized and interpreted, although such concepts clearly drive the transformation. For example, the transformation of "ASAP" to "20" or "Blue" to "10-12" or "Red" to "1 (=worst)" has no meaning to final data users, and the logic of domain experts that drives this transformation is not explicit for validation, change, or replication. These non-explicit quantification approaches can be used for nominal or ordinal (ordered) codes, but do not truly make the final data semantically comparable. For ordinal data, the ordering of data instances might be useful for some purposes, but the true patient acuity for the "worst" code at one hospital might only be equivalent to the lower ranking in another. Since quantification approaches do not explicitly target the underlying concepts, it is impossible to know if the transformed data is semantically comparable. (In other words, although "1" equals "1" in the new representation, does "ASAP" hold an equivalent meaning to "Shock" and "Red"?) Semantic comparability might indeed be achieved by this strategy, but it is not guaranteed.

The second approach to resolving different coding schemes, called translation here, deals more explicitly with the underlying concepts of each local data term or instance. The general idea is to examine local data, tease out relevant concepts, and transform the native data to a standard concept system representation. This approach is more labor intensive and conceptually challenging, but is the only approach that promises to capture the intended semantics, or true meanings, of each local term or code.

This simple example presented in Figure 1 is typical of problems encountered in controlled healthcare vocabularies with literally millions of concepts. Health care data sources utilize a host of concept systems for a variety of different purposes. In these vocabularies and coding systems, each 'value' or data instance is a terminological representation of a unique concept or group of concepts. Because the concepts represented in different concept systems are often not 1:1, or even n:1, they can be very difficult to resolve. The enormity and complexity of medical knowledge makes the assimilation of different concept systems a greater challenge than dealing with many conventional ratio measurement systems. The translation between disparate units in ratio

measurement is straightforward, since, for example, one inch always equals approximately 2.5 cm. But what is the relationship between "coughing/wheezing" and "breathing problems", or the relationship between "nasal congestion" and "runny nose/ green"? The existence of many concept systems for health care knowledge is frequently referred to as the "vocabulary problem" in the medical informatics literature, and is a major research focus for the field. [11] Standard or reference terminologies are the vehicle of choice to achieve data comparability and preserve intended semantic meaning.

2.4 Reference terminologies

A reference terminology is a terminology (i.e., set of specified concepts and inter-relationships) that functions as the standard for comparison of data from heterogeneous representations and/or collected for different purposes. Reference terminologies (to which disparate local concepts are "mapped") are critical to resolving differences in concept representation arising from heterogeneous systems, by providing a standardized list of concepts and (term) labels that have shared and consensually understood meanings across a domain or user community.[34]

The ideal reference terminology has the concept coverage and granularity to meet the intended needs and to capture local data. The fitness of a reference terminology is entirely dependent upon the purpose, and it has been said that for any purpose, no perfect terminology exists.[37] As knowledge grows and data availability increases, no one can predict all potential and future needs for a given terminology. Domain and industry vocabulary standards are never sufficient for all potential use cases, and often knowledge changes more quickly than the standards for its representation. There will always be a need to integrate data from heterogeneous database representations, and referent knowledge representations will always be called for, often customized for specific data integration projects.

In theory, a reference terminology can be created from one or all of the component terminologies, borrowed, modified, or created anew. In reality, however, the final purpose for the data integration often requires a unique reference terminology to be developed. The development and use of a reference terminology is more than a merging of all local terms, because a reference terminology needs to contain both the concept coverage and structure to satisfy final information needs, and at the same time capture *similarity* and *overlap* between the heterogeneous component concept systems. Since reference terminologies are indeed concept systems, the literature on conceptual modeling and ontology development provides good resources for design strategies.[24, 28, 31, 46] All of these sources emphasize the highly iterative development process, the reliance on domain experts, relevant literature, and data instances for development and refinement, and the importance of the purpose for which these conceptualizations are created to support. Because conceptual models and ontologies are domain dependent, there is no step-by-step methodology for their creation. While there are few standards for the design of reference terminologies per se, there are established guidelines for the evaluation of terminologies in healthcare that are generalizable in part to other domains.[36, 39, 47-50]

2.5 Mapping relationships to reference terminology concepts

In heterogeneous database integration problems, a major class of representational heterogeneities arises from differences in the underlying concept or measurement systems encoding the data in each local database, and requires transformations to a referent concept or measurement system to integrate the content in a meaningful way. "Mapping" is the asserted relation between the representation of a concept in one concept system to the most similar representation in another system [51], and is achieved by transforming (or re-coding) concepts from local data instances to a standard concept system Different native or local data representations imply variability in the nature of these data instance – reference concept mappings. In the next section, we argue that the capture of this variability, referred to here as mapping characteristics, can help preserve data granularity and intended meaning when the data is transformed to a reference standard.

5.) Case study: reference terminology development and mapping

Successful data integration solutions preserve data granularity and intended meaning of local data content while transforming them into a uniform representation for comparability. A content integration effort in the health care domain illustrates key issues in the development of a reference terminology, associated mappings, and explicit representation of characteristics of local term – referent concept mappings. The project integrated presenting complaint data from 4 different hospital emergency departments. The heterogeneous data was made homogeneous by mapping disparate data models to a referent global schema for emergency department (ED) visits and by mapping the underlying disparate terminologies (i.e., data content) to a reference terminology for presenting complaints.[4, 5]

3.1 Problem: heterogeneous data content not comparable

Data content was not comparable between the 4 study hospitals because of representational heterogeneities in the source databases. This lack of comparability was due to disparities in both data models and underlying data content. The databases from some emergency departments collected multiple presenting complaint attributes, others one. The attribute names differed: e.g., "presenting complaint", "presenting complaints", "chief complaint", or "chf_compl". More problematic was the fact that within presenting complaint-type data fields, there was enormous disparity in content, due to the variability in coding schemes (ranging in detail from 22 possible codes to 77 possible codes) as well as the presence of free-text entries. Some of the disparities in presenting complaint data content are shown in Figure 2. For example, one ED may code a visit as "difficulty breathing", while another may use "shortness of breath/wheezing". Any sensible data integration requires an understanding of whether "difficulty breathing" and "shortness of breath/wheezing", or for that matter, "respiratory problems", are the same or related, and if related, how related. For public health purposes, it was important to describe and

integrate and describe this disparate presenting complaint data across multiple hospital EDs.

Figure 2. Types of Presenting Complaint Data Content from Heterogeneous ED Databases

Emergency Department A	Emergency Department B	Emergency Department C
Respiratory		
Fever/Infection		
Arrest/Resuscitation		
	Respiratory problems	
	Fever	
	Brachycardia	
	History of: Asthma	
	Malaise: Flu-like Symptoms	
	Malaise: Irritable/Anxious	
	Cough/secretions	
		Difficulty Breathing
		Coughing/crying
		Fever vomiting cold sx
		Asthma, exacerbation
		Fussy, cough

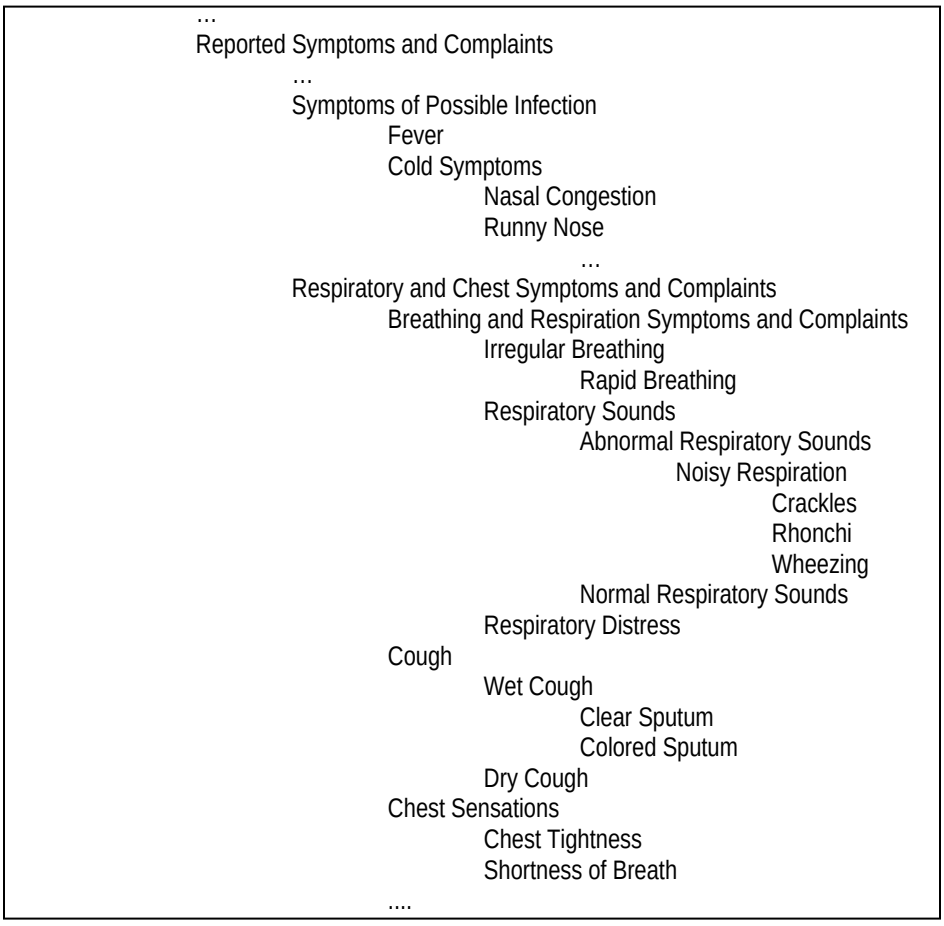
Figure 2 illustrates some typical presenting complaint data instances from 3 hospital EDs. The "units" for each coding system are concepts, which make the resolution of these coding schemes to a common representation difficult. These types of examples from Figure 2 illustrate the data instances that needed to be represented in a uniform reference terminology.

3.2 Solution: A reference terminology to capture presenting complaints

Using a global schema for ED visits, presenting complaints from several heterogeneous database representations were normalized into a single "presenting complaint" attribute in a data repository. Despite this homogeneous data structure, significant disparities in underlying concept systems remained (as shown in Figure 2), and a reference terminology was needed to provide a common representation and facilitate content integration. Existing terminologies were examined, but none had the concept specificity or semantics required to support the local data instances. The Houston Asthma Reference Terminology (HART) was developed iteratively using

domain experts, scientific literature, and actual data from pediatric ED visits from the participating Houston-area hospitals. (Figure 3)

Figure 3. HART Reference Terminology (abridged)



The structure of the HART is a (multiple) hierarchical classification. It was designed to capture the 'low-hanging fruit' of how to count, roll-up and drill-down respiratory-related presenting complaint concepts that can be used to characterize patient visits from multiple hospital emergency departments. The HART is a set of concepts (and standardized term labels) whose semantics are determined by each concept's placement in the hierarchy. The HART reference terminology provides a knowledge structure for aggregating presenting complaint data from heterogeneous ED databases, and defines the limits to which these data can be manipulated or shared. Data instances from presenting complaints were mapped (i.e., transformed) to HART concepts, yielding compiled data with a common representation.

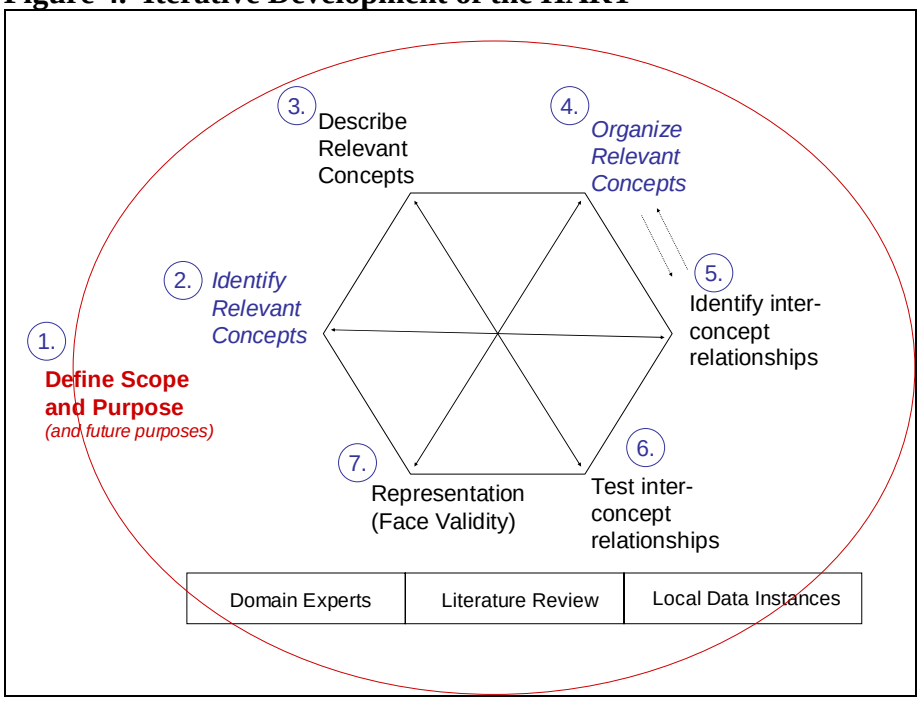
3.3 Development of the Houston Asthma Reference Terminology (HART)

The HART was developed iteratively using domain experts and scientific literature as well as actual instance data from the 4 hospital databases. The development strategy for the HART was an initial top-down conceptualization followed by many data-driven re-organizations. After identifying relevant concepts from domain experts and important scientific and professional literature, data instances from the component ED databases were examined. These data instances were mapped to concepts in the evolving reference terminology, and subsequently drove changes in the content and organization of the HART. The development of this reference terminology was neither trivial nor exact, and is described in detail in the following section. The final chapter of this paper generalizes this process for future endeavors in other domains.

Before the development of the reference terminology, representative instance data (i.e., actual data content from patient records) was filtered for respiratory diagnoses so that the data largely represented only pediatric respiratory visits for each of the EDs. This representative data (161 instances) included presenting complaints and standard visit information (e.g., date and time of service). Four domain experts (1 pediatric pulmonologist MD, 1 pediatric emergency medicine MD, 2 pediatric Emergency Department RNs) and one terminologist were consulted during the development of this reference terminology.

The HART was modeled as a hierarchy of concepts, and presented in a paper (MS Word) format for expert review and revision. The primary author developed each iterative reference terminology structure and solicited individual feedback from 4 domain experts (for content changes) and 1 terminology expert (for technical/knowledge representation changes). An iteration was defined as any addition, removal, or reorganization of relevant concepts in the HART terminology. These changes in the reference terminology were characterized by source (data-driven, expert opinion) and typology (expansion, reduction, change in inter-concept relationships). Finally, the development activities were characterized into an overview of reference terminology construction shown in Figure 4.

Figure 4. Iterative Development of the HART



The scope and intended purposes of the HART guided each development phase, and the foundation for all of the HART development activities were domain experts, relevant professional literature, and local data instances. The role of domain experts was critical. The available data and organizational restraints made it impossible to approach the actual coders for each local data instance, but domain experts in the field and familiar with each organization were consulted to understand the most likely intended semantics of each local term. The development of the HART was highly iterative, and while the key activities are discussed sequentially below, the process required many activities to be revisited and was not so simplistically sequential. The development of the HART consisted of the following activities:

1.) Define Scope and Purpose: The purpose of the reference terminology was clearly defined: *"to represent pediatric ED presenting complaints that might be relevant to or predictive of pediatric asthma."* In reality, the purpose was broader in that we wanted to retain as much detail and semantic intent as in the local databases as possible. The purpose determined what was relevant in terms of content and structure for each iteration of the HART, and was influential in every activity of the reference terminology development.

2.) Identify Relevant Concepts: Concepts included in the HART were obtained from domain experts, scientific literature, and data instances. Building upon the broad concept groupings identified by domain experts, examination of the important literature (including current asthma treatment guidelines) generated a list of potentially relevant concepts. Given that a primary purpose of the HART was to represent as much granularity as possible in the local instance data, this project placed heavy weight to the data-driven aspects of the reference terminology. Local data instances included terms

such as “cough/secretions”, “wheezing”, “tachypnea”, and “SOB”. Authors attempted to “tease apart” the underlying distinct concepts and represent them with a uniform term both recognizable and meaningful to end users, e.g., “cough”, “secretions”, “wheezing”, “fast breathing”, and “shortness of breath”. Lists of free text term entries were examined in a similar manner for underlying concepts. New local terms were mapped to the HART until saturated, meaning that no new (relevant) concepts arose. The initial listing of relevant concepts had no particular order. Subsequent exploration of the concept characteristics ultimately began to drive the structural organization of the HART.

3.) Description of HART Concepts: The listing of potentially relevant concepts facilitated “eyeballing” for like properties or characteristics. This was a highly iterative process, involving lots of pen and paper lists and scribbles, and many different hierarchical organizations were attempted. Examination of the listing of relevant concepts allowed concept “attributes” to emerge within the context of the defined purposes. These attributes were used to explore different HART organizations of groupings. New, related, and often implied, concepts that were important to represent with each data instance also were teased out, including time start/onset, when/where/who reported, and data field of origination.

4.) Organization of Important Concepts in Reference Terminology: This step was the first to try to build the HART structure, which will ultimately have implications for the quality of the transformed data, the ability of the reference terminology to represent greatest intended semantics from each source, and the capture of any similarity between component terminologies. The list of important HART concepts derived from experts, literature, and data instances, and the important concept properties and characteristics were then explored for possible modes of organization. The initial hope was that one terminology might “fit” inside another less granular terminology. The most precise local terminology (i.e., that with the greatest number of codes) was first used as the potential foundation, and mappings of concepts found in other data instances were attempted. When it was clear that several concepts from local coding systems could not be synchronized without losing data granularity from others, the idea of trying to assemble a reference terminology by overlaying local coding systems onto one another was abandoned. To visualize disparities in content and structure among the local terminologies, each local coding system was transformed into a hierarchy and viewed side-by-side. The eyeballing of the coding systems generated ideas for possible organizational structures. Each possible organization was then “tested” for the validity of its inter-concept relationships, and the inclusion of required inter-concept relationships, as described in the next step.

5.) Identify Inter-concept Relationships: The different organizational structures of the HART (i.e., each iteration) were tested by identifying relationships between multiple HART concepts. Authors tried to identify properties or characteristics that were related to (or helped describe or refine) large groups of concepts. Common underlying concepts or attributes emerged with different experimental concept groupings. For example, terms such as “wheezing”, “difficulty breathing”, “rapid breathing”, and “labored breathing” share a property “abnormal breathing” under which these concepts could all be grouped.

The identification of inter-concept relationships in each iteration of the HART organizational structure, as well as important relationships that should be represented in the HART structure, provided the basis for testing inter-concept relationships that drove subsequent iterations of the HART.

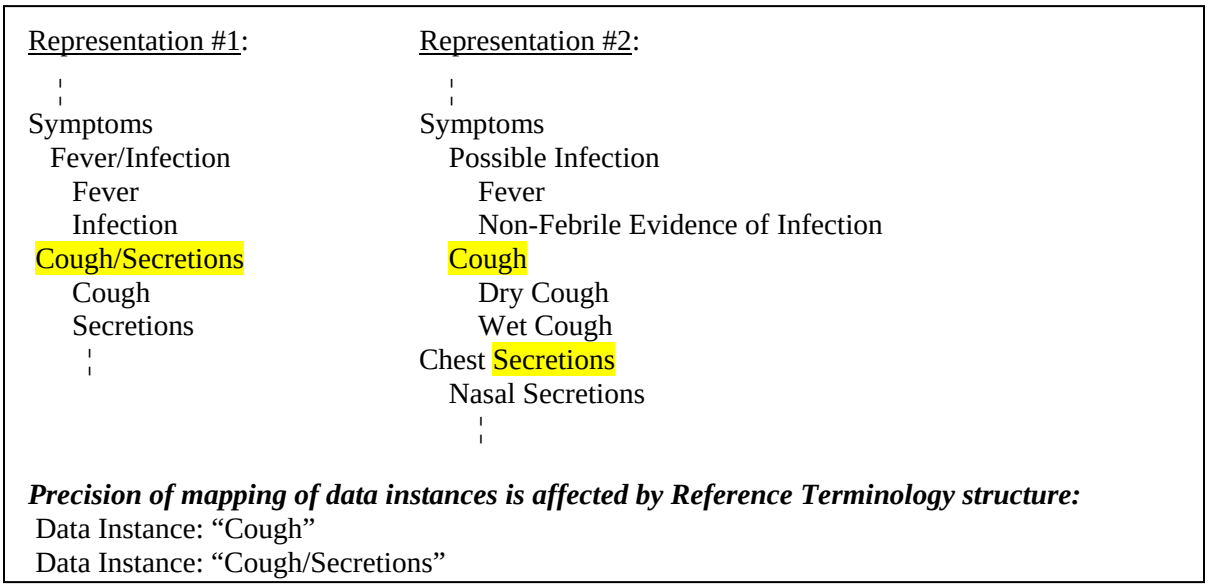
6.) Test Inter-Concept Relationships: The robustness of a terminology is determined by its usefulness or concept inclusion for local data values, and therefore the testing of inter-concept relationships was a core activity of the HART development. For example, the testing of inter-concept relationships drove questions such as: *Do all sub-groupings of abnormal breathing really share this property? Are there any important concepts related to abnormal breathing missing from this grouping?* The testing of inter-concept relationships was facilitated by concept "instances" – defined either by experts or by actual data instances. The first strategy for testing inter-concept relationships was simply to "walk the tree" and test the validity of the *is-a* relationships for the entire terminology. Next, the mappings (from local term to HART concept) were checked by domain experts and by nurses who coded the local data instances to clarify the semantic "intent" of the term. When a given instance didn't "map" to the HART, the organization of the HART was challenged, resulting in changes in content and relationships that were in turn checked by additional data instances. Characterizations for why a given instance didn't map were useful in identifying the problems with each emerging HART structure. The characterization of problems with the reference terminology structure included: a.) content problems in the HART (i.e., a concept was missing or an irrelevant concept was captured), b.) structural problems in the HART (i.e., important inter-concept relationships were missing, or incorrect inter-concept relationships were included in HART structure), and c.) problems extracting concepts encoded by local terms (e.g., multiple concepts were "lumped" into one local term, concepts were implied but not explicit in local terms, or the concepts embedded in local terms were vague or uncertain.)

7.) Representation: The selected format for the HART was determined by the intended purposes. The HART needed to explicitly represent ED presenting complaint data instances and how they could be "rolled-up" for aggregate analyses, so a hierarchy was sufficient. Basic heuristics for face validity were used to ensure the consistency of the physical representation. Terminological labels were selected to reflect the HART concepts in terms that were meaningful and familiar to potential users. Each HART concept was labeled to be identifiable independent of its placement in the hierarchy (e.g., "respiratory_and_breathing_symptoms_and_complaints" is more understandable out of the hierarchy than the label "respiratory"). Each branching level of the HART was checked for consistent levels of abstraction. Where local data instances implied a grouping or new concept, the HART was expanded to create groupings for other related or parallel concepts. For example, if a data instance drove the grouping "wet cough", then another grouping was made for "dry cough". Similarly, if terms "wheezing" and "crackles" were grouped as "abnormal respiratory sounds", then other types of abnormal respiratory sounds were identified to exhaust the abnormal respiratory sounds grouping to accommodate additional future term mappings.

3.4 Challenges for HART development

The lack of one-to-one correspondence between concepts represented by local concept systems complicated the development of the HART, and forced choices to be made about the final content and structure of the reference terminology. The presence of coded terms representing multiple or "lumped" concepts influenced the reference terminology structure. Such concepts presented opportunities for multiple different data-driven HART representations, each of which would either result in some loss of granularity for some sources, or affect the quality of the aggregated data for queries. To illustrate, two different reference terminology structures are presented as an example in Figure 5. Each potential HART structure has implications both for how the data can ultimately be used, and for the precision of the mapping of terms from local data representations.

Figure 5. Potential Reference Terminology Representations and Sample Data Instances



Some local data instances represented multiple or "lumped" concepts that each needed to be mapped to the HART. The presence of such lumped concepts raised questions for the organization of the HART. Namely, do terms such as "Cough/Secretions" or "Fever/Infection" necessarily drive similar concept grouping in the final reference terminology? While a data instance of "Cough/Secretions" maps easily to the structure in Representation #1, that representation is less robust and its uses are more limited than the second representation. For example, from Representation #1,

one cannot query for how many "coughs" (without secretions) there are across regional ED visit data. This representation is easier to construct but more difficult to query. In essence, Representation #1 illustrates a strategy of combining every possible code from the local terminologies, which fails to express any similarities between the terminologies. The chosen solution for the HART project was found by re-examining the purpose. The broader purpose of the reference terminology was to maintain as much granularity as possible, so we chose to keep concepts distinct and part of other hierarchical relationships. It was decided that dual hard-coded concepts, such as "Cough/Secretions" and "Fever/Infection" and "Cold/Infection", would not necessitate specific dual categories in the HART, because this would force a loss of data granularity for those sites that split those terms. Such disparities in different coding schemes were handled by explicitly representing the nature of the term-concept mappings, as described later. The final HART structure, therefore, includes important concepts identified by domain experts, but also captures some similarity between the local concept systems that a simple 'merging' of all component terms could not achieve.

Local terms representing multiple implied concepts were also problematic and forced choices in the structure of the final HART. The basic unit of a reference terminology should be a single concept, yet many coded instances embodied multiple concepts that were hard to tease apart. For example, the term "Flu-like Symptoms" could mean many different things, and this concept could be decomposed into all possible symptoms (e.g., chills, fever, malaise, etc.) This is challenging because the term is admittedly vague, and includes several concepts with some certainty and several others with less certainty. For the purposes of the HART, however, the relevant underlying concept was an indication of acute infection, so this concept, although vague, was not decomposed.

3.5 Characterization of HART iterations

Characterizations for iterations, mapping problems, and reference terminology changes are critical for developing a generalizable process for reference terminology development in other domains. In this project, changes in the reference terminology were characterized by source (data-driven, domain expert) and typology (HART content expansion or reduction, change in inter-concept relationships/organization). Figure 6 depicts some of the problems identified by domain experts and data instances that drove iterations of the HART, as well as their resolution.

Figure 6. Examples of Concepts Driving Changes in Content and Organization of HART

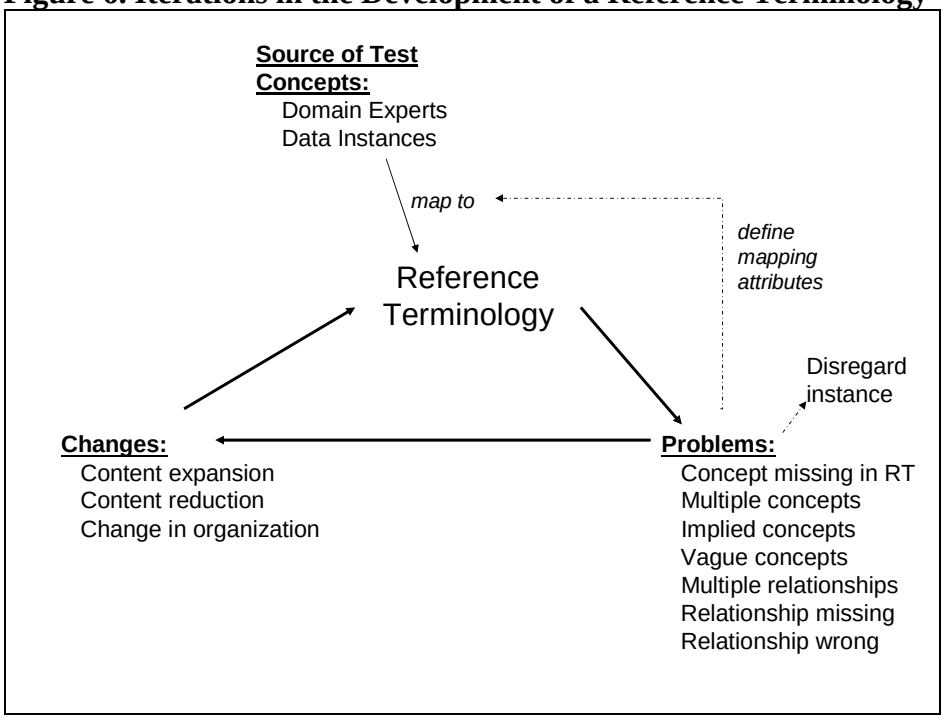
Problem	Source	Resolution	Concept	Detail
Missing Concept	Domain Expert	Concept ignored	No breath sounds	Although not present, expert pointed out that this concept could be important. Since this concept is unlikely to appear in data, the HART was not expanded.
Multiple Concepts in one Term	Data Instance: coded value	Mapping rules clarified	Wheezing/ Respiratory Distress	Respiratory Distress is not an extreme of wheezing, but is a distinct concept that can apply to any lung disease. It just refers to degree of involvement. This instance maps to 2 distinct HART concepts
Implied Concepts	Data Instance: free text entry	HART not changed, data attributes added to represent implied concepts	Asthma exacerbation	There are no consistent symptoms for an asthma exacerbation, so this concept (while vague and encompassing multiple concepts) only implies a worsening of asthma and cannot be decomposed with any certainty. This term in presenting complaint field implies a history of asthma or previous asthma diagnosis. Therefore, this term maps to HART as asthma diagnosis (patient-reported and history-of).
Vague Concept	Data Instance: free-text entry	HART content expanded	Sick	This "lay" term clearly could encompass many concepts, but none with certainty. Drove a new category of concepts under Discomfort.
Multiple inter-concept relationships	Data Instances: free text entries and coded values	HART content expanded	Pain	Pain concepts have distinct grouping organized by topography – for face validity/ completeness of RT, possible categories were exhausted. Since it was important to represent ear pain (could be a co-morbidity), earache is represented both as a Pain concept and as a Head and Neck complaint concept. (i.e., the concept is unique with several HART locations.)
Terms with Multiple Hierarchies	Data Instance: coded value	Mapping rules clarified	Congestion	Concept falls under 2 groups: possible infection and head complaint. Although not a medical term, likely that congestion could mean chest congestion. This concept maps in multiple places with lower levels of "certainty" or confidence for each mapping.
Wrong relationship	Expert	HART organization changed	Aspiration/ Choking	Mis-grouping in early HART iteration. Aspiration/Choking is actually a lower respiratory (not a digestive) symptom.
Missing Concept	Expert	HART content expanded	Peak Flow Laboratory Tests	Added concept: add peak flow to lab tests. Not data driven but important to represent.
Multiple Concepts	Data Instance: coded value	HART organization does not reflect lumped concept	Fever/ Infection	Does fever/infection dictate unique heading since it is a compound concept in some sources? These are 2 distinct concepts. To retain data granularity from other sites, concepts fever and infection were kept separate.

3.6 Logistics of HART development

Data instances and domain experts forced choices about both the content (i.e., included concepts) and the organization (i.e., how to group concepts) of the HART. Problems identified by attempts to map data instances or important concepts defined by domain experts included: missing concepts in the HART, local terms that contained multiple HART concepts, local terms that contained implied HART concepts, and local terms representing vague concepts. Domain expert review also identified problems or inconsistencies with inter-concept relationships denoted by the HART organizational structure. Comments about inter-concept relationships included both the location of a concept in the HART structure (e.g., "Aspiration/Choking" is not a 'digestive_symtom_or_complaint' but is a 'lower_respiratory_symptom_and_complaint') and important concepts with relationships to multiple groupings, (e.g., the term "congestion" is a 'symptom_of_possible_infection' and is a 'head_complaint'.) Each new HART representation is called an iteration. Each iteration was tested using data instances and domain expert reviews.

The general process for reference terminology development is illustrated in Figure 6. Each iteration resulted in the addition, removal, or movement of concepts in the emerging HART structure, or new groupings and organizational structures, and was driven by either mapping actual data instances or potential data instances identified by domain experts. Each iteration required either a change in the HART structure (in either content or organization), a decision to consider the data instance as either irrelevant to the scope of the reference terminology purpose or an anomaly, or a representation of mapping characteristics to qualify the mapping. Therefore, each iteration resulted in a resolution that could be characterized as adding/removing concepts, moving concepts, ignoring the instance, or clarifying the mapping rules and descriptions.

Figure 6. Iterations in the Development of a Reference Terminology



The development of the HART was more resource-intensive than it would appear. 161 data instances and 13 important expert-defined relationships were used to develop the HART. A total of 32 iterations were made to arrive at the final HART structure depicted in Figure 3. The development time was approximately 6 person-months, including several hours of expert review.

3.7 Characterizing the mappings to the HART

As described earlier, alternative representations of the HART have implications for both the coverage of data instances and the ultimate usefulness of the aggregated data. Some of the granularity lost in the transformation or mapping of local data values to reference terminology concepts was regained by representing the nature of these mapping relationships. Quality mapping characteristics, whose explicit representation could have potential importance in resultant applications, were identified through the development of this data integration process. These were noted by examining the types of data granularity loss incurred by different reference terminology representations, and by the use of domain experts to verify if these differences were important. With domain experts, potential characterizations of quality were identified.

The explicit representation of the precision of mapping to a reference terminology can help in retaining maximum data granularity from each local data source. For example, if one component terminology uses the term “dyspnea” to mean the referent concept “Shortness of Breath”, the fact that the two terms represent the exact same concept, i.e., they are common synonyms, implies a smaller likelihood of loss of meaning

and straying from the intent of the local term. Similarly, the mapping of "barky cough" to the referent concept of "wet cough" is far less precise, since "barky cough" can be considered a vague term. The mapping precision of lexical variants (e.g., "coughing" at the local level vs. "cough" at the referent level) falls somewhere in between. The notion of representing the precision of match is common in many web-based search applications [18, 26] and ontology integration projects.[30, 32] A distinct attribute that describes the precision of term-concept mappings can have implications for both the maintenance of the reference terminology and the quality of the aggregated data for analyses. The final representation for the precision of each local term-concept mapping in this project includes: exact term and concept, lexical variation, synonym, broader than, narrower than, related concept.

One measure of the quality of the mapping is a representation of who created or validated a given mapping assertion. Ultimately, a third party creates the term-concept mappings that transform each native term to a standard reference terminology concept in the final data repository, and representing the source of these mapping assertions can improve the quality of the transformed data. For example, the fact that a physician, or more specifically, a pulmonologist, reviewed a term-concept mapping was an important quality attribute for this project. Depending upon the application or use case, it might be of greater importance to know that the triage nurse who actually coded a term validated its semantic intent. The best-suited mapping experts differ for the problem or information-type being examined. Explicitly representing this basic measure of quality of each term-concept mapping assertion adds power to the aggregated data, facilitating the potential quantification of certainty about the match. The final representation for quality developed for this project captures who asserted the local term-HART concept mapping (medical expert from ED where instance originated, medical expert from other ED, nurse coder from ED where instance originated, nurse coder from other ED, or health informatics developer). Another measure of quality that was considered here was a certainty factor, similar to peer reviewers' comments on their certainly that a work should be accepted. This should be correlated with precision, above. Although quantification of this quality attribute is a future work, its explicit representation is a start.

Most database integration efforts transform data to a common representation making the user unaware of any disparities in the native representations. Processes that address these differences in local term-referent concept mappings can enhance the end users' understanding, querying, and use of aggregated data. This is particularly critical when heterogeneous data integration efforts entail the transformation of data to a new conceptual reference model. Making these mapping characteristics explicit has implications both for the evaluation and maintenance of the reference terminology and for utility of the transformed data in computational or statistical analyses.

6.) A Process for concept integration

Based upon the experiences in the data content integration effort described thus far, this section describes a generalized process for the integration of heterogeneous

concept systems to a uniform representation (i.e., reference terminology), with explicit relationships for the representation of quality that can preserve the data granularity and semantic meaning. This process is part of a larger process for heterogeneous database integration[5], and includes both the development and use of a reference terminology as a standard concept system for heterogeneous data content. Despite the potential variability in both project requirements and local data sources across other applications, successful efforts for integrating heterogeneous concept systems include the broad steps, identified below:

1.) ***Define purpose, information needs, and process needs*.** This step is critical. The purpose and information needs for the aggregated data dictate the level of detail and organizational structure required for the reference terminology. Information needs are best represented by creating typical "use cases" that illustrate type, detail, and applications that the aggregated homogeneous data needs to support. The purpose should be defined by a representative sample of potential application users, terminologists, and domain experts.

2.) ***Examine concept systems, data instances, and data collection context from local data sources***. The knowledge representation systems for each data attribute should be explored. What data types do they represent? If ordinal, what are the concepts that each code represents? What are the similarities in rankings across ordinal coding schemes? If concept systems (i.e., coding systems or terminologies) represent data content, what are the definitions, characteristics and properties of each concept? What are the similarities in content and semantic relationships across all concept systems? Can local database administrators and data collection persons identify any implied concepts for given codes? What other aspects of context can help identify the intended meaning of the data values? Do the contexts/settings/quality of data collection vary by site? What are the relationships between concept systems? Are there equivalencies? Can one concept system fit inside another?

3.) ***Define reference terminologies.*** Each data element in the global schema should be considered for a reference terminology that reflects expression and context of local data and encompasses the representation needs for all potential use-cases. Existing terminologies should be considered for re-use because the development of a reference terminology is a resource-intensive task. The level of disparity in concept systems across local databases and the information needs for the integrated data (both organizational and granularity) determine whether the reference terminologies can be borrowed from other sources, or must be created via top-down or bottom-up approach. Key activities for reference terminology development include:

-clearly define representation and expression requirements
-list relevant concepts
-describe relevant concepts
-identify important inter-concept relationships
-organize relevant concepts

-test inter-concept relationships
-choose appropriate representation

4.) ***Map terms (and the concepts they represent) expressed in local data sources to the closest concept in the appropriate reference terminology.*** The focus of this mapping needs to be on the underlying concept(s) or intended concept(s) expressed as a term in local databases. Semantic intent is determined from either observing or questioning both developers and representative coders for each local data source, as well as domain experts. The exploration of the context of data collection, as well as the roles and training of local data entry persons from each data source, can assist in understanding semantic intent.

5.) ***Represent precision and quality of mapping of local data instances to concepts in reference terminology.*** Each data value in each concept system should be mapped to the most appropriate and closest concept in the reference terminology. Variability in concept systems from the heterogeneous sources implies a loss of data granularity from some sources, and this loss of granularity should be represented in the final application as appropriate to the needs of the integrated data. The identification of potential context items that impact the quality of match is facilitated by domain experts and designers of local data sources. An appropriate representation for the variable mapping (precision and quality) of local terms to referent concepts should be developed.

The steps outlined above are highly iterative, and certain steps will entail returning to previous activities. The first step of identifying the needs and purposes for the data integration should be re-visited at every step, and is the lens through which evaluation of the process and resultant homogeneous data is ultimately determined.

While the process of validly aggregating data from heterogeneous databases into a common representation is highly dependent upon the domain, purpose, and the nature of the concept systems encoding local source data, the overall goals and measures of success are similar. The ultimate goal for content integration is to create “comparable” data in an organization and representation suitable for a given purpose. Given the final information needs, the most successful strategy should retain as much granularity from as many sources as possible. The evaluation of success is addressed by examining the structure of the reference terminology, and the relationships between the local data values to the reference terminology.

Some broad questions can be asked to evaluate the success of the process:

2.) Have the purposes for the integration been clearly identified and specific use cases created? Do all stakeholders agree?

3.) Does the reference terminology include all of the concepts and semantics required to meet the use cases?

4.) Does the reference terminology have the granularity and detail needed to support the final use cases? Does the reference terminology selected for each construct limit the loss of data granularity from each local data source? Is the loss of data

granularity acceptable to the final purpose and use cases for the aggregated homogeneous data? Does the reference terminology capture all important similarities in concepts and relationships across the component concept systems?

5.) Is there an explicit representation for the quality of the mapping of local data values to each reference terminology? Is this representation useful or meaningful to domain experts in rating the quality of the data transformations?

6.) Is there an explicit representation for the precision of the mapping of local data values to each reference terminology? Is this representation useful in explaining the variability of mappings and loss of granularity?

7.) Was the context of data collection observed for disparities in the operationalization of data definitions? Were all aspects of context explored for their potential role in determining semantic intent of local data values? Have domain experts identified important elements of context that might impact the quality of the final data?

In general, the evaluation of the process for content integration is guided by the intended purposes of the integration project. The generalized process and evaluation criteria described above ensure a systematic approach for the examination of underlying semantics of local concept systems and reference terminologies, and the evolving relationships between the two. The intended purposes and final information needs drive each iteration of this process as well as define its completion and success.

5.) Conclusions and future work

The fundamental principle of the resolution of disparate concept systems is to capture underlying semantics. First, a true understanding of the nature of important concepts is needed, which is facilitated by exploration of domain experts' conceptualization of the domain and context of data collection. Different conceptualizations and representations are possible, and each can have implications for the integrated data.

As a knowledge representation, a reference terminology is a conceptualization of a domain for a given purpose. There are multiple possible representations and an ideal representation must include all needed concepts and be robust enough to represent real data instances. Often, the nature of the heterogeneous concept systems forces developers to loose some data granularity from some sources. The ideal integration solution includes a reference terminology, supported by transformation procedures that explicitly represent relevant mapping characteristics, and therefore minimizes the loss of data granularity and preserves the intended semantics from each local source.

This process for valid content integration from heterogeneous concept systems is one part of a larger database integration solution, and addresses a problem that is often not fully realized. Similar endeavors will need to determine if and where reference terminologies are needed, adopt, modify, or develop the reference terminologies,

integrate them into a global schema to eliminate structural database heterogeneities, and finally consider mechanisms to keep the reference terminologies updated.

The contribution of this work is a process for constructing and implementing a reference terminology to integrate data encoded by heterogeneous concept systems. The characterizations of problems, iterations, and changes in the evolving reference terminology present a framework for further content integration efforts. The representation of local term – reference terminology mapping characteristics enhances the utilization of a reference terminology to preserve local data granularity and semantics as well as to capture similarities that exist between local concept systems. While the exact quality and precision representations used here may not meet the needs of other applications, the theory of identifying and quantifying these mapping characteristics should enhance the validity of a number of database integration efforts in a number of domains. The generic notion of representing variable local term – referent concept mapping relationships is a novel approach to retaining local data semantics and capturing similarity across heterogeneous data representations. While the characterization of these mapping characteristics might not meet all uses, it is our feeling that this is a critical piece of developing intelligent data content integration applications solutions. Current database integration efforts transform data to a common representation, but the user is often unaware of any disparities in the native representations. Processes that address these differences in local term-referent concept mappings ultimately enhance the end users' understanding, querying, and use of aggregated data.

Merging heterogeneous data representations includes resolving representational differences on many levels, some of which are not fully explored. The separation of content integration from broader data integration activities emphasizes the critical role of these activities in facilitating integration of heterogeneous data into a useful and uniform representation that allows comparability while preserving the intent and detail of the original sources. It is hoped that the process presented here will motivate further research on the refinement, elaboration, and evaluation of the development and use of reference terminologies for heterogeneous database content integration.

AKNOWLEDGEMENTS:

The authors wish to thank Dr. Charles Macias from the Department of Pediatric Emergency Medicine at the Baylor College of Medicine and Dr. Marianna Sockrider from the Department of Pediatric Pulmonology at the Baylor College of Medicine for their expertise, assistance and enthusiasm throughout this project.

This research was facilitated by the Robert Wood Johnson Foundation: Managing Pediatric Asthma: Emergency Department Demonstration Program; Pediatric Texas Emergency Department Asthma Surveillance (TEDAS).

Rachel Richesson is funded by National Library of Medicine Fellowship in Applied Informatics #1 F32 LM07188-01A1.

References

1. Richesson, R.L., et al., *Creating Homogeneous Data from Heterogeneous Representations: A Process for Heterogeneous Database Integration.* Submitted to: Data and Knowledge Engineering, 2003.
2. Richesson, R.L., et al., *Foundations for Heterogeneous Database Integration: A Framework to Identify Representational Heterogeneities.* Submitted to: ACM, 2003.
3. Lee, M.L. and R. Ramakrishnan, *Integration of Disparate Information Sources: A Short Survey.* ACM Multimedia, 1999.
4. Mitra, P., G. Wiederhold , and M. Kersten. *A Graph-Oriented Model for Articulation of Ontology Interdependencies*. in *Proceedings of Conference on Extending Database Technology*. 2000. Konstanz, Germany.
5. Raschid, L. and Y.H. Chang, *Interoperable Query Processing From Object To Relational Schemas Based On A Parameterized Canonical Representation.* International Journal of Cooperative Information Systems, 1995.
6. Sheth, A.P. and J.A. Larson, *Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases.* ACM Computing Surveys, 1990. **22**(3): p. 183-236.
7. ISO, *2000 Terminology Work - Vocabulary - Part 1: Theory and Application (Final Draft International Standard)*. 2000, International Organization for Standardization.
8. Sujansky, W., *Methodological Review. Heterogeneous Database Integration in Biomedicine.* Journal of Biomedical Informatics, 2001. **34**: p. 285-298.
9. Bakken, S., et al., *Toward Vocabulary Domain Specifications for Health Level 7-Coded Data Elements.* Journal of the American Medical Informatics Association, 2000. **7**(4): p. 333-342.
10. Spackman, K.A., K.E. Campbell, and R.A. Côté, *SNOMED RT: A Reference Terminology for Health Care.* Journal of the American Medical Informatics Association, 1997. **4(Symposium Supplement)**: p. 640-644.
11. Neces, R., et al., *Enabling Technology for Knowledge Sharing.* AI Magazine, 1991. **12**(3).
12. Sugumaran, V. and V.C. Storey, *Ontologies for Conceptual Modeling: Their Creation, Use, and Management.* Data & Knowledge Engineering, 2002. **42**: p. 251-271.
13. McGuinness, D.L., *Conceptual Modeling for Distributed Ontology Environments.* Proceedings of the Eighth International Conference on Conceptual Structures Logical, Linguistic, and Computational Issues (ICCS 2000), 2000(August 14-18, 2000).
14. Musen, M.A., *Ontology-Oriented Design and Programming*, in *Knowledge Engineering and Agent Technology*, J. Cuena, et al., Editors. 2002, IOS Press: Amsterdam.
15. Burgun, A., et al., *Issues in the Design of Medical Ontologies Used for Knowedge Sharing.* Journal of Medical Systems, 2001. **25**(2): p. 95-108.

16. Chute, G.C., S.P. Cohn, and J.R. Campbell, *A Framework for Comprehensive Health Terminology Systems in the United States.* Journal of the American Medical Informatics Association, 1998. **5**(6): p. 503-510.
17. Campbell, J.R., et al., *Phase II Evaluation Of Clinical Coding Schemes: Completeness, Taxonomy, Mapping, Definitions, And Clarity.* Journal of the American Medical Informatics Association, 1997. **4**(3): p. 238-251.
18. Bakken, S., et al., *An Evaluation of the Usefulness of Two Terminology Models for Integrating Nursing Diagnosis Concepts into SNOMED Clinical Terms.* International Journal of Medical Informatics, 2002. **68**: p. 71-77.
19. Shiffman, R.N., et al., *Information Technology for Children's Health and Health Care: Report on the Information Technology in Children's Health Care Expert Meeting, September 21-22, 2000.* JAMIA, 2001. **8**(6): p. 546-551.
20. Cimino, J., *Desiderata for Controlled Medical Vocabularies in the Twenty-First Century.* Methods of Information in Medicine, 1998.
21. Elkin, P.L., et al., *Guideline and Quality Indicators for Development, Purchase and Use of Controlled Health Vocabularies.* International Journal of Medical Informatics, 2002. **68**(1-3): p. 175-186.
22. ISO, *Working Document: Health Informatics - Vocabulary on Terminological Systems*. 2000, International Organization for Standardization.
23. Sycara, K., M. Klusch, and J. Lu. *Matchmaking Among Heterogeneous Agents on the Internet*. in *AI Spring Symposium on Artificial Agents on Cyberspace*. 1999.
24. Nodine, M., J. Fowler, and B. Perry. *Active Information Gathering in InfoSleuth*. in *International Symposium on Cooperative Database Systems for Advanced Application*. 1999.
25. McGuinness, D.L., et al., *The Chimaera Ontology Environment.* American Association for Artificial Intelligence, 2000.
26. Mena, E., et al., *Observer: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies.* International Journal Distributed and Parallel Databases, 1998.

Obtaining Comparable Presenting Complaint Data From Heterogeneous Emergency Department Databases*

Rachel L. Richesson, PhD, MPH[1,*], James P. Turley, RN, PhD[1], Kathy A. Johnson-Throop, PhD[1], Christoph Eick, PhD[2], Mark S. Tuttle, FACMI[3]
[1]University of Texas Health Science Center at Houston, School of Health Information Sciences, [2]University of Houston, Department of Computer Science, [3]Apelon, Inc.

*[SUBMITTED TO JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION, AUGUST 2003]

***Corresponding Author:**
Rachel Richesson, PhD, MPH
University of Texas Health Science Center at Houston
School of Health Information Sciences
7000 Fannin, Suite 600
Houston, TX 77030 USA
713-500-3456, 713-500-3915 (FAX)
Rachel.L.Richesson@uth.tmc.edu

ABSTRACT

Integrating data from heterogeneous databases into a homogeneous representation presents both a conceptual and practical challenge, and is necessary to achieve the comparability required for any kind of aggregation or manipulation of the underlying data. One strategy to achieve this is to use a reference model which provides the standard for how data from different knowledge representations can be integrated in a meaningful way. An example of this strategy is illustrated in a demonstration project to make heterogeneous emergency department (ED) presenting complaint data “comparable” and therefore enable subsequent integration and aggregation. A general process for achieving a homogeneous data representation is applied to a set of heterogeneous ED databases, resulting in the Houston Asthma Reference Terminology (HART), and associated "maps", with which locally-coded pediatric ED presenting complaints can be analyzed. The HART reference model is empowered by a global data schema for ED visits, which includes quality and precision information that enhance the meaning of the aggregated data, empowering end users with better-informed queries for subsequent analyses. This solution of a global ED visit schema and supporting reference terminology for presenting complaints resulted from a general process that can be repeated in other heterogeneous database integration problems.

Keywords: Heterogeneous database integration; content integration; reference terminology development; data quality; public health informatics

I.) BACKGROUND

A.) Problem: Regional Presenting Complaints not Comparable

Asthma is currently the most common chronic disease in children.[52] For many reasons related to chronic disease labeling, pediatric asthma may not be diagnosed. In the acute emergency department (ED) setting, any number of pediatric respiratory illnesses may mimic asthma, further complicating its diagnosis. Often, EDs, rather than primary care, are used as a source of care for episodic and acute asthma exacerbations.[53] For this reason, comprehensive community surveillance should include many settings, including EDs. Because of lack of specificity of ICD-9 discharge diagnosis data and documented under-diagnosis of asthma by physicians [54], an alternate source of electronically-captured ED visit information is desirable for identifying potential pediatric asthma in the ED setting. The Texas Emergency Department Asthma Surveillance (TEDAS) project, funded by the Robert Wood Johnson Foundation, is a consortium of researchers and physicians from 4 Houston-area EDs interested in developing effective ED-based asthma surveillance methods and health interventions.

Currently, counting and examining pediatric respiratory or asthma ED visits across a region is difficult because the presenting complaint data characterizing these visits are not "comparable" across heterogeneous hospital databases. This lack of comparability is due to representational disparities in data models and data content across multiple hospitals, and results in a need for solutions to integrate data into a common representation. One activity of the TEDAS project was to develop the Houston Asthma Reference Terminology (HART), which provides a uniform representation with which to compare

(and subsequently count and aggregate) presenting complaint data from pediatric respiratory visits at multiple EDs throughout the region. This paper describes the development and use of the HART as part of a general and repeatable process for creating a comparable and homogeneous representation from heterogeneous database representations.

B.) Representational Differences Across ED Databases

Many representational differences across ED databases are due to differences in local database schema and the underlying concept representation (e.g., coding or terminology) or measurement systems. To illustrate, Figure 1 shows sample data instances from the database schema of 3 different EDs used in the TEDAS data integration process.

Figure 1. Sample Data Instances Using Different Database Schema for Emergency Department Data Capture (abridged)

(note: data from same 2 patients represented in 3 different ways)

Emergency Department A

Patient #	Date of Service	Age	Chief Complaint	Acuity
123456	10-24-01	12	Cough/Fever/Malaise	Mild
234567	10-24-01	3	Respiratory Distress	Severe

Emergency Department B

Medical Record #	DOS	Time of Service	Age	Acuity
123456	10-24-01	0300	12	Mild
234567	10-24-01	1400	3	Severe

Medical Record #	Presenting Complaints
123456	Cough
123456	Fever
123456	Malaise
234567	Respiratory Distress

Emergency Department C

Visit #	Social Security #	Date/Time of Service	Date of Birth	Description	Value
888	123456	10-24-01 3:00am	11-1-1991	Presenting Complaint	Cough
888	123456	10-24-01 3:00am	11-1-1991	Presenting Complaint	Fever
888	123456	10-24-01 3:00am	11-1-1991	Presenting Complaint	Malaise
888	123456	10-24-01 3:00am	11-1-1991	Acuity	Mild
999	234567	10-24-01 2:00pm	3-6-2000	Presenting Complaint	Respiratory Distress
999	234567	10-24-01 2:00pm	3-6-2000	Acuity	Severe

Despite the broad similarity in workflow and data capture needs across EDs, a current lack of standards results in an enormous variety of database implementations. The data instances from Emergency Department A's database schema design, for example, represent all information in one table that includes data attributes for a visit identifier, date of service, age of patient, presenting complaint (labeled "chief complaint"), and acuity. Data instances from Emergency Department B contain a unique patient medical

record number, date of service, time of service and patient age in one table that is related to a separate table with multiple instances of presenting complaint values. Those data from Emergency Department C's database present yet another valid information structure, with one table containing a visit number, patient social security number, a combined attribute for date and time of service, date of birth, and a description attribute (presenting complaints or acuity) with a value in the value attribute. The differences in data representation shown in Figure 1 are due to the specific information needs of each organization, and to the developers' conceptualization of these needs. Such schematic differences result in naming, formatting, and structural differences that must be resolved to unify the data into a common, or homogeneous, representation.

In addition to heterogeneities introduced by different database schema, significant differences between ED databases are introduced by different representations or concept systems used to encode presenting complaint values (i.e., data content) at each site. The ED databases integrated by this project used different, locally-created coding schemes, ranging in detail from 22 possible codes to 77 possible codes. One ED may code a visit as "difficulty breathing", while another may use "shortness of breath/wheezing". Any scalable public health application needs to know whether "difficulty breathing" and "shortness of breath/wheezing", or for that matter, "respiratory problems", are the same or related, and if related, how related. Figure 2 illustrates some of the types of presenting complaint data represented in 3 different ED concept systems.

Figure 2. Selected Presenting Complaint-Type Data Values from Heterogeneous Emergency Department Databases

Emergency Department A	Emergency Department B	Emergency Department C
Respiratory		
Fever/Infection		
Arrest/Resuscitation		
	Respiratory problems	
	History of: Asthma	
	Malaise: Flu-like Symptoms	
	Malaise: Irritable/Anxious	
	Cough/secretions	
		Difficulty Breathing
		Coughing/crying
		Fever vomiting cold sx
		Asthma, exacerbation
		Croupy cough
		Crying for 4 hrs

Emergency Departments A and B use coded attributes, and C uses unstructured data to represent reported presenting complaints. Each represents a concept system, with the underlying unit of 'measures' being concepts. In order to answer such questions as *"How many kids presented to regional EDs with coughs? Breathing problems? How many potential asthmatics presented to regional EDs for care?"*, a common representation or terminology is needed. This means that concepts must be transformed, or mapped, from one representation to another. Mapping is the relation between the representation of a concept in one terminological or concept system to the most similar representation in another concept system. [35] The lack of 1:1 correspondence between the concepts underlying the terms in these disparate knowledge representation systems makes their resolution to a common representation, or reference terminology, a challenge. Possible approaches include merging all of the local codes and terms into a comprehensive coding system, choosing one concept system as a standard and mapping all other systems to it,

using an outside concept system as a reference model, or creating a new reference model that captures the semantics of each local concept system. For this application, we chose the last approach since it shows the most potential to retain important semantics from each local concept system yet capture similarities that might exist between them.

C.) Heterogeneous Databases in Health Care

The problems described above are typical of heterogeneous databases in healthcare. The resolution of multiple representational disparities to achieve homogeneous aggregate data is important for a variety of applications, including improved healthcare, decision-making, outcomes research, evaluation, public health surveillance, and bio-terrorism preparedness. Heterogeneous databases are defined as separate autonomous databases, independently created for unique purposes, with substantial differences in database schema. [1] The integration of heterogeneous database schema and associated concept and measurement systems requires the use of one or more referents, or standards, to which the component data structures or data values are mapped. *Global database schema* are referent models that guide the integration of heterogeneous database schema into a uniform representation, and *reference terminologies* are referent models that guide the integration of heterogeneous concept systems (e.g., terminologies, coding schemes) into a uniform data representation. In heterogeneous data integration projects, global database schemas and reference terminologies address the issues of *schema integration* and *content integration*, respectively, and together provide a uniform representation of data from heterogeneous database sources. The goal for integrating heterogeneous databases is to achieve comparable data with a homogeneous representation from different source

representations; success is determined by retaining as much granularity (i.e., depth and detail) and intended meaning as possible from each source. [45]

D.) Overcoming Semantic Heterogeneity

While the structural and content disparities across heterogeneous databases mentioned thus far are difficult to resolve, a more burdensome class of *semantic heterogeneities* can be introduced by both heterogeneities in database schema and underlying data representations. Broadly, semantic heterogeneity occurs when there is a disagreement about meaning, interpretation, or intended use of same or related data[8], and arises from different definitions of data attributes, differences in coding precision of the data values across multiple databases [1], or context [9]. Semantic heterogeneity in part refers to the fact that data in different systems may be subject to different interpretations, even when data types, labels, and general schemas are identical. [10] For example, if presenting complaint-type information collected from one ED included those reported by the patient, but those collected from another ED routinely collected nurse observations in addition to patient-reported complaints, there would be a semantic mis-match between the operational definitions of apparently similar attributes. Semantic heterogeneity is difficult to precisely define, identify, and classify,[1] yet there is common consensus that semantic heterogeneity is the class of heterogeneity that threatens multiple data conflicts, and the most problematic aspect of heterogeneous database integration efforts. [8] [10]

Examining the intended semantic meaning of each data structure or value and mapping it to the closest concept in the referent models (i.e., global schema and reference terminologies), as is the approach in this demonstration, preserves semantic intent and

can therefore reduce semantic heterogeneity. However, the mapping of heterogeneously represented data to uniform reference models implies that the local – standard relationships differ by source. For the duration of this paper, we will refer to local term–reference terminology relationships as 'term-concept mappings'. To retain data granularity and intended semantic meaning, detail about these term-concept mapping relationships must be identified and explicitly represented. This detail is a measure of mapping quality.

1.) Quality of mapping

The key to reducing semantic heterogeneity and preserving data granularity is to capture the intended semantics of each item. Ultimately, a third party retrospectively creates the term-concept mappings that transform each native term to a standard reference terminology concept. In heterogeneous database content integration, quality can be defined as the truthfulness in asserted term-concept mapping transformations. The importance of explicitly representing who created or validated a given mapping assertion can be an important characteristic to represent when looking at the aggregated data. Further, characteristics about the reliability, competence, and training of the reviewer can be captured, as well as a confidence rating that different reviewers assign to term-concept mappings. Explicitly representing these basic measures of quality of the term-concept mapping assertion can add power to the transformed data, facilitating the potential quantification of certainty about the match.

The explicit representation of the precision of mapping can also help in retaining maximum data granularity from each local data source, and improve the quality of the

transformed data for secondary analyses. The frequent lack of 1:1 correspondence of concepts represented in disparate concept systems denotes variability in precision of the term-concept mappings across sources. For example, if one local terminology uses the term “dyspnea” to mean the referent concept “Shortness of Breath”, the fact that the two terms represent the same concept, i.e., they are common synonyms, could mean a smaller likelihood of loss of meaning or straying from intent of the local term. Similarly, the mapping of “barky cough” to the referent concept of “wet cough” is less precise, since “barky cough” can be considered a vague term. The mapping precision of lexical variants (e.g., “coughing” at the local level vs. “cough” at the referent level) falls somewhere in between. The precision variability is therefore a data attribute that can preserve some data granularity. A distinct attribute that describes the precision of term-concept mappings can have implications for both the maintenance of the reference terminology and the quality of the transformed data for analyses. A representation of precision of match is common in many web-based search applications [18, 26], ontology integration projects[30, 32], as well as the National Library of Medicine’s MeSH and UMLS concept systems.

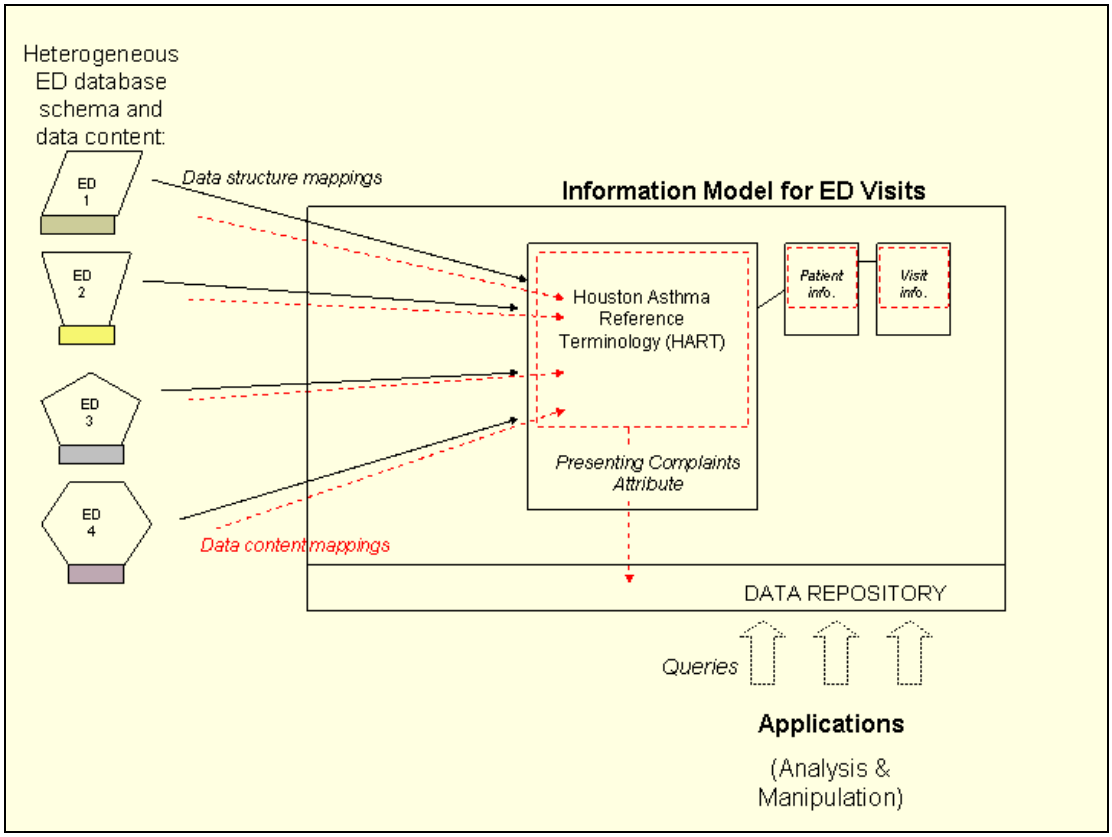
The many representational and semantic heterogeneities encountered in this project to integrate ED presenting complaint data from heterogeneous databases are typical of problems encountered in any health care data integration project. The resolution of these disparities to a uniform representation for comparability requires the use of referent standards and associated mappings from each local data structure or instance.[1, 10, 55] Heterogeneous database integration requires uniform representation in both database schema and underlying concept and measurement systems. Successful solutions preserve

as much granularity and detail as possible from the local systems.[45] The reduction of semantic heterogeneity and the preservation of data granularity from each source can be facilitated by examining the quality and precision of the term-concept mappings. A process for the integration of heterogeneous healthcare databases was applied to aggregate heterogeneous ED presenting complaint data into a uniform representation while minimizing the loss of data granularity and preserving local data semantics. The result of this process is a data repository and associated database schema that uniformly represent ED presenting complaint data from heterogeneous sources, mapped to a reference terminology, with explicit characterizations of the mapping relationships between native and final data representations.

II.) METHODS

The purpose of this project was to develop a homogeneous data representation from heterogeneous ED presenting complaint data. A homogeneous data representation was achieved by creating a global database schema for ED visits, to which presenting complaint data structures (data attributes or instances) from component databases were mapped. A reference terminology was created to represent instances of presenting complaint-type data uniformly, within the context of the global schema or information model. The need to maintain data granularity and preserve semantic intent motivated the representation of mapping characteristics (i.e., quality, precision) in the final data repository schema. The global database schema is a relational data model, implemented in an Access™ database.

Figure 3. HART Project Overview



The project overview in Figure 3 illustrates the use of a uniform representation for heterogeneous ED presenting complaint data. Local data structures (data attributes and instances), along with proposed data needs (i.e., purpose and use cases), drove the development of the HART reference terminology, which provides a uniform representation for presenting complaint data within the context of a global schema (or information model) for ED visits. The general process used to create the homogeneous data representation is presented below and discussed more thoroughly in [5]. This process addresses schematic heterogeneity, heterogeneity from underlying concept or measurement systems, and semantic heterogeneities that can result from both.

1.) Define purpose, information needs, and process needs. This step is critical and should guide all choices to be made in the design of homogeneous representation standards and use. The purpose and information needs for the aggregated data dictate the level of detail and organizational structure required for the conceptual referents. Information needs are best represented by creating typical “use cases” that illustrate type, detail, and applications that the aggregated homogeneous data needs to support. Process needs (e.g., access, timing, data availability) drive the logistical procedures of the heterogeneous database integration. The purpose should be mutually defined and endorsed by a representative sample of potential application users, database integrators, and domain experts.

The model of database federation (i.e., the general approach for accessing and integrating data) is based upon the needs defined above, and determines the practical implementation of the database integration effort. The access permissions and anticipated needs for updated or current data determine whether a query-modification or data-translation approach should be taken. In general, needs for current and frequently updated data are best satisfied with a query-modification approach, whereas data-translation is suitable for periodic data needs.[8] Also, data sources with highly disparate concept systems requiring one or more reference terminologies will need a data-translation step. Specific guidance for selecting from different models of database federation can be found elsewhere.[1]

2.) Examine data structures, concept and measurement systems, and data collection context from local data sources. Each local data source must be explored to determine

the semantic content. This examination can be *bottom-up*, meaning each local database is examined attribute by attribute, or *top-down*, meaning relevant constructs are identified from a conceptual model and the corresponding data structures or attributes are sought in each of the local database schema. Regardless of the approach, all relevant data attributes should be reviewed and synonymy in attribute names noted. In addition, the operationalized data definitions for each data attribute should be identified. The concept and measurement systems for each data attribute should be explored. This preliminary, almost qualitative, analysis of source database schema and underlying concept and measurement systems should identify each data attribute of interest in the final project and attempt to define initial equivalency relationships across databases. The level of disparities observed in structure, naming, format, data definitions, and concept or measurement systems encoding each data structure will dictate the best approach for defining the global schema. Accordingly, the activities associated with this step are a prerequisite for step 3, defining the global schema.

3.) Define global schema. Depending upon the levels of disparities between the local databases and the overall project purposes, a schema integration approach (data-driven) or a top-down schema creation approach could be used. Regardless of the method, the global schema should define the constructs and relationships needed for the application, at both the level of granularity needed and with the terminology (attribute labels) familiar to the domain. The potential disparities in data attribute definitions across source databases and context of data collection that impacts these final operational definitions, as well as important quality attributes identified by domain experts, should be represented

in the global schema. Available domain ontologies (including conceptualizations of process and work-flow) should be searched for relevance and used as a resource to guide the development and/or refinement of the global schema. Ideal global schema should maintain relationships to the local data schema that allow the traceability of the native data context.

4.) Define reference terminologies and measurement systems. All concept and measurement systems must have a standard representation. Each data structure in the global schema should be considered for a reference terminology that reflects expression and context of local data and encompasses representation needs for real use-cases. The level of disparity in concept systems across local databases and the information needs for the aggregated data (both organization and granularity of content) determine whether these reference terminologies should be borrowed from other sources, or created via top-down or bottom-up approach. One reference terminology might capture the concepts represented in multiple data attributes of the global schema. The use of reference terminologies to achieve comparability from heterogeneous data can be termed content integration, and a generic process for this is described in [2].

5.) Map data structures expressed in local data sources to the closest construct in the global schema. This step generally can be thought of as reconciling local database attributes to corresponding attributes in the global schema, but can also involve moving data from instance level to attribute level to table level. Regardless, the meaning or concept class represented by each structure should be the focus of this activity. To

preserve semantic intent from each local source, the focus goes beyond the data definitions of each database to include interviewing designers and users of each local database. Questions to be asked should include: *What is the meaning of this attribute? How is the content or value selected? Do all users agree? Does the context of the data collection influence the meaning of the data values? If so, how?* This examination of constructs at the local level might drive changes in the global schema. The ultimate information needs and purpose should guide the mapping of relevant local data structures to the appropriate structure in the global schema.

6.) Map relevant concepts that are implied but not explicit in local data models to the global schema. The global schema should identify constructs or structures to compare to the local schema. Where constructs are missing, but are implied or can be derived from local sources, they must be imputed into the global schema. For example, if the construct of who recorded a particular local data attribute is important to the quality of the final data semantics, this concept should be included in the global schema and the values imputed appropriately by local users and domain experts. The project logistics determine how the imputation process should best occur. This can be achieved by a "blanket" imputation (i.e., all values for 'reported by' are the same for a given source) or by selective value-based imputation (i.e., presenting complaint in hospital A is a 'diagnosis' structure if it contains the term 'asthma'.) A representation for missing concepts that cannot be implied or derived should be included in the global schema. The analysis of disparities in data collection context, and review by domain experts and end-users,

facilitates the identification, representation, and mapping of implied concepts from local databases to the global schema.

7.) Map terms (and the concepts they represent) expressed in local data sources to the closest concept in the appropriate reference terminology. Again, the focus of this mapping needs to be on the underlying concept or intended concept expressed as a term in local databases. Semantic intent is determined by observing representative coders for each local data source, as well as questioning coders, local database developers, and domain experts. The exploration of the context of data collection, as well as the roles and training and objectives of local data entry persons at each level can assist in understanding semantic intent.

8.) Characterize the quality of mapping of local data instances to concepts in reference terminology. The identification of potential context items that impact the quality of match is facilitated by domain experts and designers of local data sources, and also by a domain ontology, if available. An appropriate representation for the certainty of mapping of data structures to the global schema, as well as who asserted each mapping, should be developed. Any variability in the data definitions of constructs (e.g., one data attribute definition is operationalized differently than a corresponding construct in another local database schema) should be explicitly represented relative to each mapping relationship. Domain experts and end users should specify representations for quality that are useful and meaningful to the final applications.

9.) Characterize precision of mapping of local data instances to concepts in reference terminology. Similarly, the intended meaning of each term in each concept system should be mapped to the most appropriate and closest concept in the reference terminology, and characterizations of the local term-referent concept mappings should be represented. Variability in concept systems from the heterogeneous sources implies a loss of data granularity from some sources, and this loss of precision should be represented in the final global schema as appropriate to the needs of the final compiled data. Any variability in the data definitions of constructs (e.g., one data attribute definition is broader in scope or more inclusive than that of the corresponding construct in the reference terminology) should be explicitly represented relative to each mapping assertion.

Three key activities of the above data integration process are highlighted in the Methods and Results sections: the development of the global schema, the development of the reference terminology, and the representation of mapping characteristics.

A.) Global Schema for Emergency Department Visits

The local data schema from the 4 different EDs, plus domain experts (nurses and physicians) practicing in ED settings, were used to develop a global conceptualization of the ED visit process. Each local schema was systematically examined to determine all collected attributes (names and definitions) as well as the data collection or process activities corresponding to each attribute. The basic model of ED care was confirmed by asking several experts about the process from both the patient's point of view and from a data collection perspective. The local data schema were examined from each source to see if the data attributes collected matched the conceptualization of ED visits given by experts. The expert conceptualizations and local data models were consolidated to define

a broad data model for ED visits, which served as the global schema for the data repository. The main purpose of this global schema was to transform different representations (table, attribute, and instance level) of presenting complaint information from each local data schema to a common attribute. The global schema can be thought of as an information model that governs the assimilation of data elements to form a logical patient record.[22] This process, while important, is addressed in current research and automated solutions exist. The greater challenge for this integration demonstration is the content integration using a reference terminology, as described in the next section.

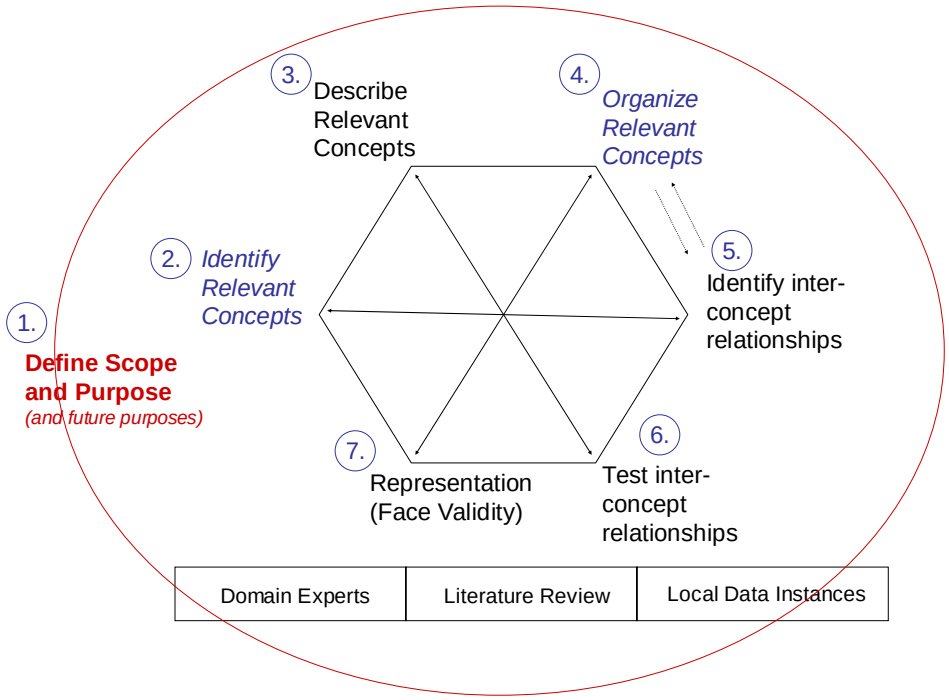
B.) Developing a Reference Terminology to Compare Presenting Complaints from Heterogeneous Representations

Once heterogeneous ED respiratory-related presenting complaints were normalized into a common attribute, a common representation (i.e., reference terminology) was needed for the data instance values. Existing terminologies were examined, but none had the specificity or organization required to describe presenting complaint data related to asthma. The Houston Asthma Reference Terminology (HART) was developed iteratively using domain experts, scientific literature, and actual data from pediatric ED visits from the participating Houston-area hospitals. Before the development of the reference terminology, representative instance data was filtered for respiratory diagnoses so that the data largely represented only pediatric respiratory visits for each of the 4 EDs. This representative data (161 instances) included presenting complaints and standard visit information (e.g., date and time of service). Four domain experts (1 pediatric pulmonologist MD, 1 pediatric emergency medicine MD, 2 pediatric Emergency

Department RNs) and one terminologist were consulted during the development of this reference terminology.

The HART was modeled as a hierarchy of concepts, and presented in a paper format for expert review and revision. After its development, it was represented in an Access™ database, with one table listing all HART concepts, and a separate related table of specified parent-child relationships. The primary author developed each iterative reference terminology structure and solicited feedback from 4 domain experts (for content changes) and 1 terminology expert (for technical/knowledge representation changes). An iteration was defined as any addition, removal, or reorganization of relevant concepts in the HART terminology. These changes in the reference terminology were characterized by source (data-driven, expert opinion) and typology (expansion, reduction, inter-concept relationships). [2]

Figure 4. Iterative Development of the HART



The development strategy for the HART was an initial top-down conceptualization followed by many data-driven reorganizations. After identifying relevant concepts from domain experts and important scientific and professional literature, data values from the component ED databases were examined. These data instances were mapped to concepts in the evolving reference terminology, and subsequently drove changes in the content and organization of the HART. The development was highly iterative but included the main activities shown in Figure 4. The scope and intended purposes of the HART guided each development phase, and the foundation for all of the HART development activities were domain exerts, relevant professional literature, and local data instances.

C.) Mapping Characteristics

The important role of mapping characteristics (quality and precision) in the preservation of data granularity and intended meaning of the local data terms was illuminated by the process of building the HART. The variability in local term-HART concept mappings sometimes demonstrated a loss of granularity or intended meaning from some source representations, and some of this lost data granularity and semantics could be captured in new data attributes representing the quality and precision of the match. For example, if local terms “dry cough” and “cough” are both mapped to the concept “cough” in a reference terminology, the transformed data would appear similar (i.e., 2 instances of cough) but the distinctions between the local terms would be lost. If this loss of granularity was acceptable to the final application (e.g., all we want to do is count coughs) then the final reference terminology representation perform adequately. The capture of the different relationships between term-concept mappings for both local terms, however, can allow some data granularity and potential meaning to be retained. Noting that the local term “dry cough” is more granular or specific than the reference concept “cough” and that the local term “cough” represents the same concept as reference concept “cough” can facilitate understanding of the relationship that exists between the two local terms. Most database integration efforts transform data to a common representation making the user unaware of any disparities in the native representations. Processes that address these precision differences in local term-referent concept mappings can enhance the end users’ understanding, querying, and use of aggregated data. Making these mapping characteristics explicit has implications both for the evaluation and maintenance of the reference terminology and for utility of the transformed data in computational or statistical analyses.

III.) RESULTS

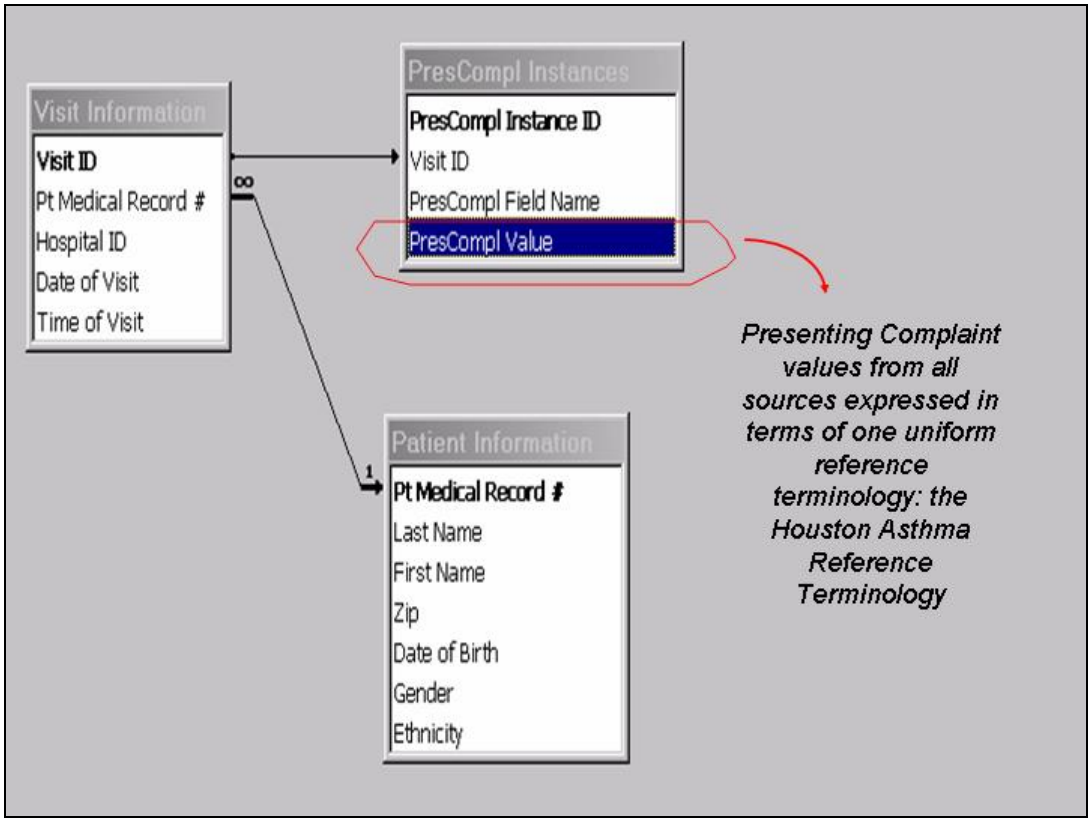
The application of the process above resulted in a global schema for ED visits and the HART reference terminology, to which disparate data representations from component databases were mapped, as well as explicit representation of the characterization of those mappings. Essentially, the global schema, or information model is the result of the structural or schema integration, and the HART is the result of the content integration.

A.) Global Data Schema for Emergency Department Presenting Complaints

It was important to resolve as many physical differences as possible before dealing with terminology issues across presenting complaint-type data, which were the greater challenge in this project. The process of assimilating heterogeneous database schema to a global model developed quickly. A partial global data schema for ED visits, highlighting presenting complaints, is illustrated in Figure 5. The model was created by looking at the types of data attributes available in each of the 4 local databases, and from the domain experts' characterization of the ED visit process. The process model used to support development of this global schema is that patients (who have more or less permanent characteristics) present with complaints, have an acuity/severity, receive one or more diagnoses, and leave the ED with a final disposition status (e.g., home, admitted to hospital, etc.). The resulting global schema builds from this conceptualization, and divides the information from each ED visit into 2 main tables: 1.) patient information (demographics, identifier) linked to 2.) visit-specific information (date/time of visit, hospital of visit, visit identifier). Each visit is related to separate tables for presenting complaints, and diagnosis, acuity, and disposition (not shown). For any given ED visit,

one or more presenting complaint instances can occur in the global schema. This data model "normalizes" presenting complaint data from the local data attributes ("Chief Complaint", "Presenting Complaint #1", "PresCompl", etc.), creating a common data attribute ("PresCompl Value") for all presenting complaint data instances.

Figure 5. Partial Global Schema for ED Visits, Highlighting Presenting Complaint Data



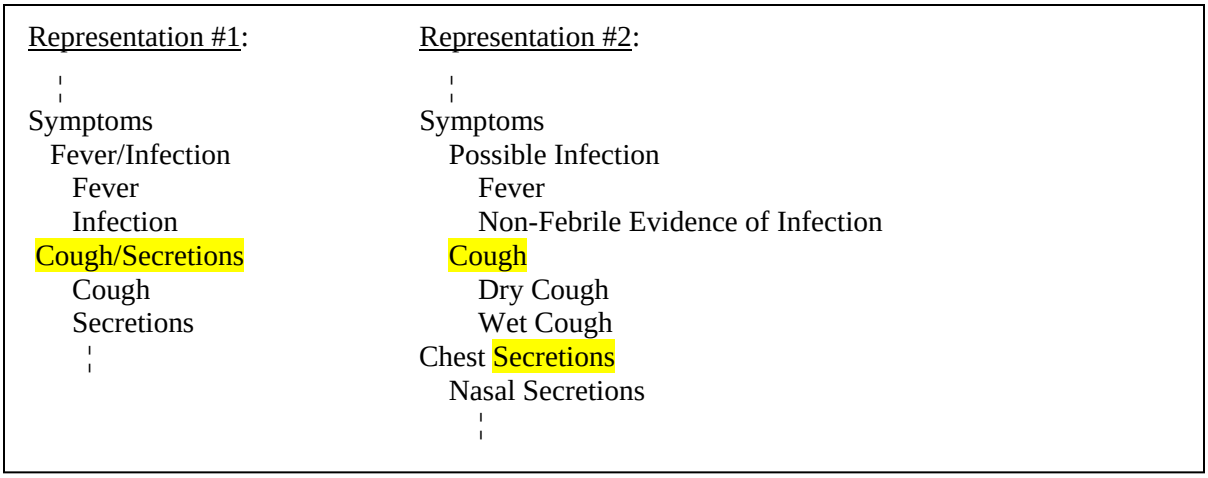
The global schema puts all complaint instances into a single complaint attribute ("PresCompl Value"), all diagnoses into a single diagnosis attribute (not shown), all acuity ratings into a single acuity attribute (not shown), etc. The PresCompl Value represents the original data values from each component database, and are not

“comparable” despite now being represented uniformly in a single data attribute. Therefore, a uniform representation, or reference terminology was needed.

B.) Development of the Houston Area Reference Terminology (HART)

The development of the HART was influenced both by domain experts and by actual data instances. Concepts and inter-concept relationships introduced by both influences the final HART structure, by driving changes in both concept and structure. [2] The lack of one-to-one correspondence between concepts represented by local coding or measurement systems complicated the development of the HART. The presence of coded terms representing multiple or “lumped” concepts influenced the reference terminology structure. Such concepts presented opportunities for multiple data-driven HART representations, each of which could result in some loss of granularity for some sources, or affect the quality of the aggregated data for queries. Two different reference terminology structures are presented as an example in Figure 6. Each potential HART structure has implications both for how the data can ultimately be used, and for the precision of the mapping of terms from local data representations.

Figure 6. Potential Reference Terminology Representations (simplified) and Sample Data Instances



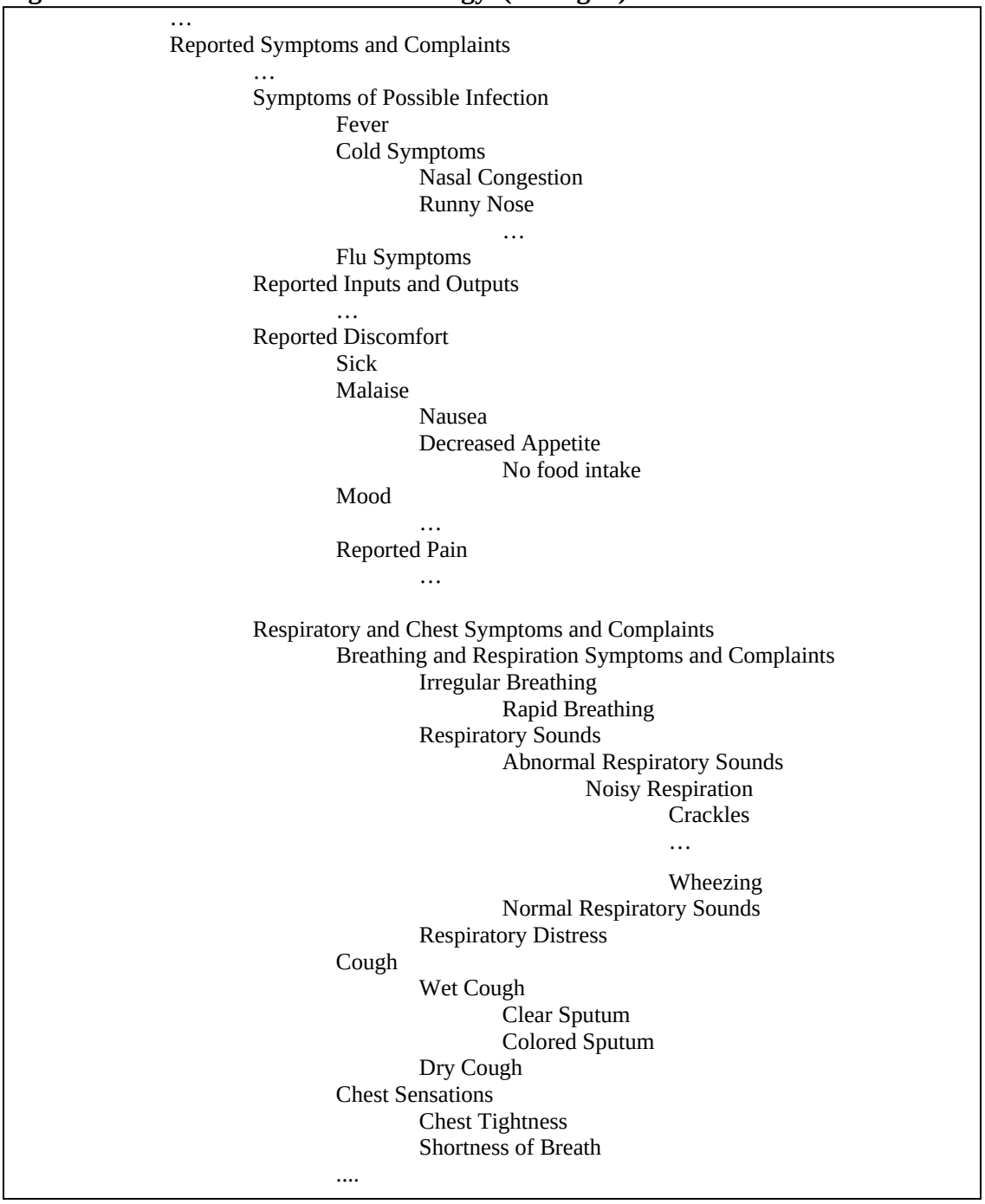
Precision of mapping of data instances can vary by choice of Reference Terminology structure:
Data Instance: "Cough"
Data Instance: "Cough/Secretions"

Local data instances representing multiple or "lumped" concepts drove questions about the final HART structure. Namely, do terms such as "Cough/Secretions" or "Fever/Infection" necessarily drive similar concept grouping in the final reference terminology? While a data instance of "Cough/Secretions" maps easily to the structure in Representation #1, that representation is less robust and its uses are more limited than the second representation. For example, from Representation #1, one cannot query for how many "coughs" (without secretions) there are across regional ED visit data. The chosen solution for the HART project was found by re-examining the purpose. The broader purpose of the reference terminology was to maintain as much granularity as possible, so we chose to keep concepts distinct and part of other hierarchical relationships. It was decided that dual hard-coded concepts, such as "Cough/Secretions" and "Fever/Infection" and "Cold/Infection" would not necessitate specific dual categories in the HART, because this would force a loss of granularity for those sites that split those terms. Such

disparities in different coding schemes were handled by explicitly representing the nature of the mappings in the final global schema, as described later.

Other terms representing multiple implied concepts were also problematic and forced choices in the structure of the final HART. The basic unit of a terminology should be a single concept[39], yet many local codes embodied multiple concepts that were hard to tease apart. For example, the term “Flu-like Symptoms” could mean many different things, and this concept could be decomposed into all possible symptoms (e.g., chills, fever, malaise, etc.) This is challenging because the term is admittedly vague, and includes several concepts with some certainty and several others with less certainty. For the purposes of the HART, however, the relevant underlying concept was an indication of acute infection, so this concept was not decomposed. A partial illustration of the final HART organizational hierarchy is shown in Figure 7.

Figure 7. HART Reference Terminology (abridged)



The development of the HART was more resource-intensive than it would appear. Data instances and domain experts made forced choices about both the content (i.e., included concepts) and the organization (i.e., how to group concepts). The problems identified by

trying to map data instances or important concepts defined by domain experts included: missing concepts in the HART (e.g., a nurse identified “no breath sounds” as a potential presenting complaint), local terms that contained multiple HART concepts (e.g., “Wheezing/Respiratory Distress”), local terms that contained implied HART concepts (e.g., the presenting complaint “Asthma Exacerbation” implies a worsening of symptoms and a likely previous diagnosis), and local terms representing vague concepts (e.g., “sick”). Domain expert review also identified problems or inconsistencies with inter-concept relationships denoted by the HART organizational structure. Comments about inter-concept relationships included both the location of a concept in the HART structure (e.g., “Aspiration/Choking” is not a ‘digestive_symtom_or_complaint’ but is a ‘lower_respiratory_symptom_and_complaint’) and important concepts with relationships to multiple groupings, (e.g., the term “congestion” is a ‘symptom_of_possible_infection’ and is a ‘head_complaint’.) New HART representations were attempted to accommodate data instances and expert reviews.

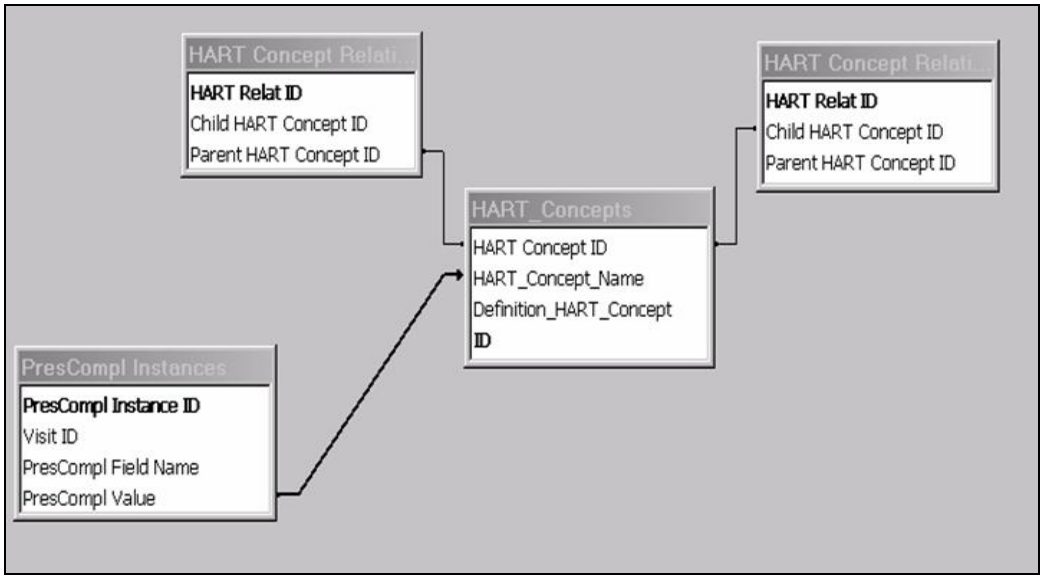
Each new HART representation was called an iteration. Each iteration then had to be checked against data instances and domain experts reviews. Each iteration resulted the addition, removal, or movement of concepts in the emerging HART structure, or new groupings and organizational structures, and was driven by either mapping actual data instances or potential data instances identified by domain experts. Each iteration required either a change in the HART structure (in either content or organization), a decision to consider the data instance as either irrelevant to the scope of the reference terminology purpose or an anomaly, or a representation of mapping characteristics to qualify the

mapping. Therefore, each iteration resulted in a resolution that could be characterized as adding/removing concepts, moving concepts, ignoring the instance, or clarifying the mapping rules and descriptions. 161 data instances and 13 important expert-defined relationships were used to develop the HART. A total of 32 iterations were made to arrive at the final HART structure depicted in Figure 7. The development time was approximately 6 person-months, plus several hours of expert review.

Like any reference terminology, the HART provides the knowledge structure for aggregating presenting complaint data from heterogeneous ED databases, and defines the limits to which data can be manipulated or shared. Data instances from presenting complaints were mapped to the final HART terminology (i.e., transformed), yielding aggregated data with a common representation.

Once the HART content and structure were finalized (Figure 7), a formal representation was needed to facilitate the mapping of local presenting complaint values to the reference terminology within the data repository. The HART was added to the data repository as 2 tables. The structure of the HART is hierarchical, and is represented in the repository as a table of concepts, twice joined to a table of parent-child concept relationships. (Figure 8) While the difficulties representing hierarchical data structures in a relational model have been observed[56, 57], the relational model was better suited for the intended uses of aggregated ED presenting complaint data.

Figure 8. Uniform Representation of Presenting Complaints to HART Concepts



Each presenting complaint value (PresCompl Value) in the repository schema maps to one or more concepts in the HART. Therefore, all presenting complaint concepts are represented homogeneously according to the HART reference terminology, as shown in Figure 8. The choices made in the development of the final HART structure affected the nature of local term-HART concept mappings. The preservation of data granularity and semantic intent was attained by adding attributes to describe the nature of each term-concept mapping. The specific attributes and values for these term-concept mapping characteristics emerged from examining how local data instances actually mapped to each evolving HART structure, as well as domain experts' opinions on what was important.

C.) Mapping Characteristics

The quality and precision for each term-concept mapping affect the data granularity and intended meaning in the global repository, and the examination of context facilitated the process for achieving homogeneous data representation while preserving data granularity and semantics. The construction of the reference terminology and the ultimate purpose of the project facilitated representations of these constructs.

1.) Representation of Quality

Quality of mapping attributes that were relevant to the TEDAS project included who asserted the local term-HART concept mapping (medical expert from ED where instance originated, medical expert from other ED, nurse coder from ED where instance originated, nurse coder from other ED, or health informatics developer) and on what date. For this application, all mappings were ultimately checked by a domain expert, but it was envisioned that the scalability of this process would benefit from a representation of who makes each mapping assertion. Mapping assertions are dependent upon local data values and the organization from which they originate. Therefore, these concepts cannot be inserted globally, but must be part of some sort of programming logic within the data repository. The usage and meaning of certain terms can vary by location and so mappings made by experts familiar with the local data coding processes carry more weight in this representation system. The explicit representation of the quality of mapping did not add much weight to this aggregated presenting complaint data achieved from this project, but this step of the process might be more important to obtaining valid assimilation of data from heterogeneous sources in other projects.

Precision describes accuracy or degree of refinement in local term-concept mappings, and can be considered a surrogate measure of quality. Characterization of this precision emerged from the development of the HART. The final representation for the precision of each local term-concept mapping includes: Exact term, lexical variation, synonym, broader than, narrower than, related concept. The representation of the mapping precision in the final data repository allows the end-user to understand the compiled data that they are querying. In addition to allowing an end user to make better-informed queries of aggregated data, this type of representation has implications for the maintenance of the reference terminology. Searching the relationships in the final data repository, one can find how many term-concept mappings have a 'narrower than' relation, for example, and can identify where additional, more granular concepts might be included in the reference terminology.

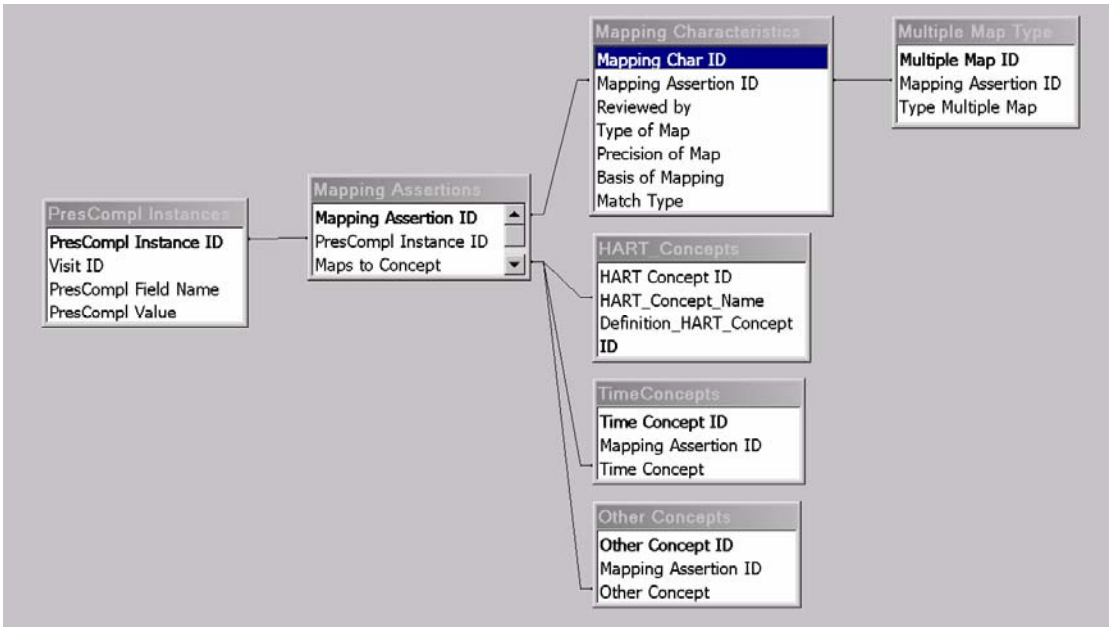
Another measure of precision that maintains semantic intent of a term in this integration, is the description of term-concept mappings that include multiple HART concepts. These types of mappings, called *multiple maps*, were characterized as conjoint, exclusive disjoint, and non-exclusive disjoint. This mapping attribute was developed to overcome choices made in the HART construction by data instances encompassing multiple or lumped concepts. For example, the intended meaning of the local coded term "Cough/Secretions" could be cough, secretions (wet cough) or both. The final representation of the HART dissected this lumped concept into two separate concepts, so the local term maps twice. But the relationship is a non-exclusive disjoint, meaning the

certainty is less than 100% for each term-concept mapping. A free-text entry of "cough/secretions", on the other hand, likely means that both concepts were intended by the coder, and the multiple mapping assertions are represented as conjoint mean that the term maps to both HART concepts with certainty. The enumeration of the certainty of mapping for these types of multiple maps adds a level of quality to the aggregated data. Inclusion of these attributes allows the final data repository to be queried to prove such relationships (e.g., all conjoint mappings can be queried and determined that they arise from free-text data structures).

The final data repository schema relates each local presenting complaint term into discrete concepts that are mapped to the HART reference terminology. Each mapping assertion is related to a table of mapping quality characteristics, by the combination of hospital, attribute name, and local term-HART concept assertion. The quality and precision mapping characteristics described above were added to the data model as 1:1 relationships with each assertion (each term-concept mapping has exactly one quality and one precision attribute).

In the final global schema, one-to-many relationships were created from presenting complaint attribute name and values to qualifying attributes (quality and precision) for the data element or the mapping.

Figure 9. Data Model for Final Data Repository



As shown in Figure 9, the final data model builds from the asserted term-concept mapping architecture shown in Figure 8, and includes explicit representation of the native data models, native data representation, and mapping (quality) characteristics. This process results in a data repository data schema that is more expressive than other heterogeneous database integration models. Typically, database integration efforts strive to create seemingly comparable data. All of the data would appear the same to the user, but often the comparability is really not there. Because of the variability in source representations, and purposes for aggregated data, and many possible representations of a reference terminology, an exact recipe for the final reference models cannot be prescribed. However, the basic process used here provides a final data schema that contains valuable qualifying data about the aggregated data. The steps in this process force the developer to examine the intended meaning and context of each local value and the final schema includes an explicit representation for these data attributes. In addition, this process recognizes that different reference terminology representations are possible,

each impacting the transformed homogeneous data in varying ways. The explicit representation of variability in local term-concept mappings can recapture meaning and precision that is often lost in traditional database integration approaches.

IV.) DISCUSSION

This work uses a generalized process for the resolution of heterogeneous databases with heterogeneous measurement systems to a common representation. The global data schema and a supporting reference terminology provide a uniform representation for heterogeneous data, and the incorporation of mapping characteristics in this schema retains semantic intent and preserves some data granularity. The use of this process, and the resulting data schema, allows public health researchers to compare and aggregate Presenting Complaint data from multiple EDs across a region to assess trends, identify community health problems and monitor ED utilization.

The development of any reference terminology is an imperfect science, heavily dictated by the application domain and user requirements, and therefore it is impossible to prescribe a step-by-step methodology. There are many possible ways to organize and represent a reference terminology, and ultimately, the best representation depends upon the purpose and use requirements. Major activities in the creation of a reference terminology are: defining purpose and scope, identifying relevant concepts, describing relevant concepts, organizing relevant concepts, identifying inter-concept relationships, testing inter-concept relationships, and representation. Literature on the development of conceptual models and ontologies supports these design stage conceptualizations.[25, 30,

58] The direction of modeling approach (top-down vs. bottom-up) can have enormous implications for the final reference terminology structure. Often, it is superior to identify broad organizational groupings before examining the underlying data. In cases or domains where the data contents are largely unknown to experts, or where preservation of local data granularity is a major objective, as was the case with this application, a more data-driven start can be warranted. Ultimately, the approach and final reference terminology content, organization, and format are left to the designer and are driven by the functional requirements, but will have implications for the transformed data. The approaches for use of a reference terminology described in this paper increase the likelihood of informed use of the data and should be valuable in a variety of informatics applications.

The immediate purpose for this endeavor was to grab the low-hanging fruit of a simple terminology to uniformly and extensionally define ED presenting complaints relevant to potential pediatric asthma in the ED. The HART provides the framework to make heterogeneously represented data comparable. This comparability is often taken for granted but uses are infinite. A repository of comparable presenting complaint data could be mined to identify groups of symptoms or complaints for a patient. The reference terminology serves this purpose, and is the instrument with which to count pediatric respiratory presenting complaints across multiple EDs in a region. The result of any database integration project is queryable data that appears the same to the end user. This process and resultant data repository schema allow the end user to identify the qualifications of the apparently comparable data from heterogeneous sources and

therefore make better-informed queries. Statisticians and database designers, and intelligent applications all will benefit from the explicit representation of how things are mapped, e.g., allowing “certainty” factors to be calculated for given mappings. The overlaying of local data schemes onto HART can illustrate the data capture limitations for any specific ED, and might stimulate sites to expand or revise data collection.

To address the absence of an existing terminology with the content and semantics required for this integration, the HART reference terminology was created and is in essence the result of the content integration, which expresses similarity in intended meaning of the instances from heterogeneous concept systems. Rather than adding to the “vocabulary problem”, this process creates some order where there once was chaos, and allows data content integration that was previously impossible to achieve. Further, this process has potential to “assimilate” existing data sources into a common representation that captures a consensus of concept representation practice and needs across a range of real-world data collections, and offer this aggregate knowledge as an extension to existing knowledge representations, such as SNOMED.

The final organization of the HART and the mapping characteristics forced the developer to examine the intended semantic meaning of terms at the local level. The examination of intended semantics was facilitated by examining context on many levels. This process has demonstrated that development choices in the reference terminology affect the quality and precision of the transformed data in a variable manner. While the exact codings used here may not meet the needs of other applications, the theory of identifying

and quantifying these mapping characteristics could be of value to a number of database integration efforts. This process and resultant data schema minimize the loss of data granularity and semantic meaning when transforming data to a homogeneous representation.

ACKNOWLEDGEMENTS

The authors wish to thank Dr. Charles Macias from the Department of Pediatric Emergency Medicine at the Baylor College of Medicine and Dr. Marianna Sockrider from the Department of Pediatric Pulmonology at the Baylor College of Medicine for their expertise, assistance and enthusiasm throughout this project.

This research was facilitated by the Robert Wood Johnson Foundation: Managing Pediatric Asthma: Emergency Department Demonstration Program; Pediatric Texas Emergency Department Asthma Surveillance (TEDAS).

Rachel Richesson is funded by National Library of Medicine Fellowship in Applied Informatics #1 F32 LM07188-01A1.

References

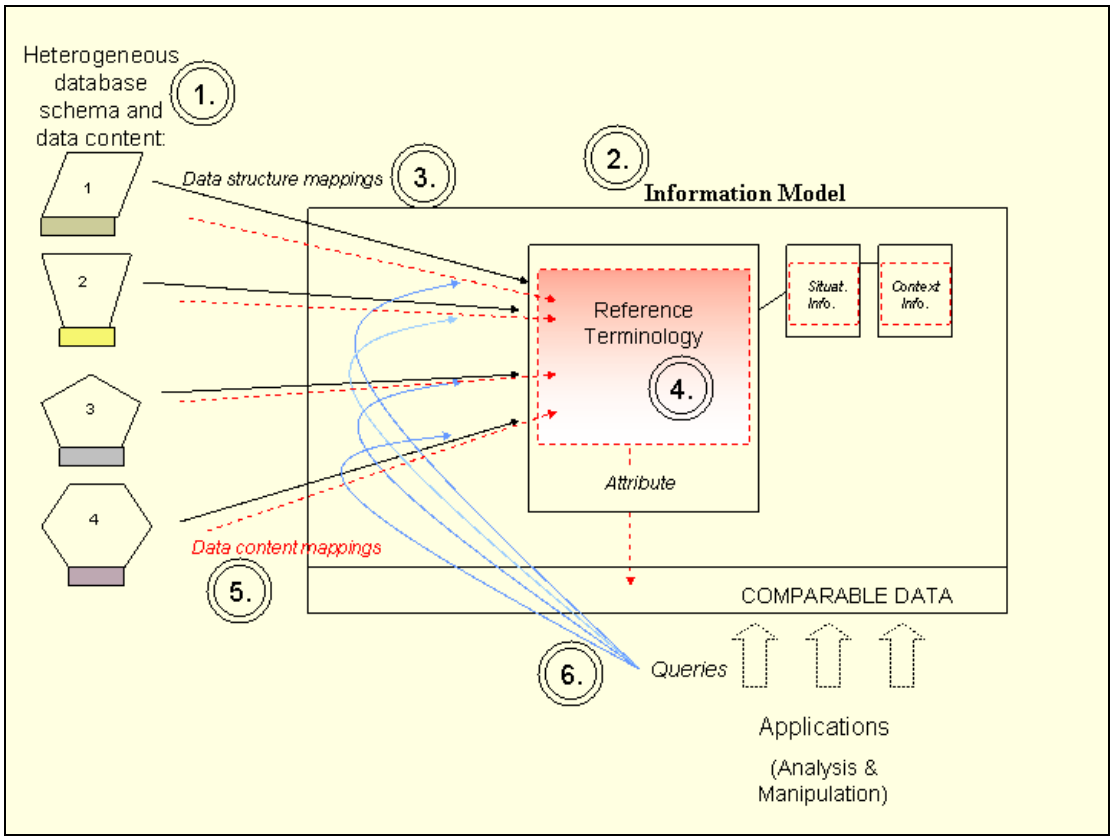
1. CDC. Disabilities Among Children Aged <17 Years - United States, 1991-1992. Morbidity and Mortality Weekly Report 1995;44:609-613.
2. Wasilewski Y, Clark NM, Evans D, Levison MJ, Levin B, Mellins RB. Factors Associated with Emergency Department Visits by Children with Asthma: Implications for Health Education. American Journal of Public Health 1996;86(10):1410-5.
3. Frank PI, Frank TL, Cropper J, Hirsch S, McL Niven R, Hannaford P, et al. The Use of a Screening Questionnaire to Identify Children with Likely Asthma. British Journal of General Practice 2001;51:117-120.
4. ISO. 2000 Terminology Work - Vocabulary - Part 1: Theory and Application (Final Draft International Standard): International Organization for Standardization; 2000 2000. Report No.: ISO/FDIS 1087-1.
5. Sheth AP, Larson JA. Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. ACM Computing Surveys 1990;22(3):183-236.
6. Richesson RL, Turley JP, Johnson KA, Tuttle MS, Eick C. Heterogeneous Database Integration: Resolving Representational and Semantic Heterogeneity to Achieve Homogeneous Aggregate Data. manuscript in progress 2003.
7. Sujansky W. Methodological Review. Heterogeneous Database Integration in Biomedicine. Journal of Biomedical Informatics 2001;34:285-298.
8. Dampney CNG, Pegler G, Johnson M. Harmonising Health Information Models - A Critical Analysis of Current Practice. In: Ninth National Health Informatics Conference; 2001; Canberra ACT, Australia; 2001.
9. Lee ML, Ramakrishnan R. Integration of Disparate Information Sources: A Short Survey. ACM Multimedia 1999.
10. Sycara K, Klusch M, Lu J. Matchmaking Among Heterogeneous Agents on the Internet. In: AI Spring Symposium on Artificial Agents on Cyberspace; 1999 March 1999; 1999.
11. Nodine M, Fowler J, Perry B. Active Information Gathering in InfoSleuth. In: International Symposium on Cooperative Database Systems for Advanced Application; 1999 1999; 1999.
12. McGuinness DL, Fikes R, Rice J, Wilder S. The Chimaera Ontology Environment. American Association for Artificial Intelligence 2000.
13. Mena E, Kashyap V, Sheth A, Illarramendi A. Observer: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies. International Journal Distributed and Parallel Databases 1998.
14. Fang D, Hammer J, McLeod D. The Identification and Resolution of Semantic Heterogeneity in Multidatabase Systems. In: Proceedings of International Workshop on Interoperability in Multidatabase Systems; 1991 April 1991; Kyoto, Japan; 1991.
15. Richesson RL, Turley JP, Johnson-Throop KA, Eick C, Tuttle MS. Creating Homogeneous Data from Heterogeneous Representations: A Process for Heterogeneous Database Integration. Submitted to: Data and Knowledge Engineering 2003.
16. Richesson RL, Turley JP, Johnson-Throop KA, Eick C, Tuttle MS. Development and Use of a Reference Terminology to Maintain Data Granularity and Semantics in the

Integration of Heterogeneous Data Content. Submitted to: Data and Knowledge Engineering 2003.
17. Huff S, Carter J. A Characterization of Terminology Models, Clinical Templates, Message Models, and Other Kinds of Clinical Information Models. In: AMIA Symposium; 2000; 2000.
18. Cimino J. Desiderata for Controlled Medical Vocabularies in the Twenty-First Century. Methods of Information in Medicine 1998.
19. Schadow G, Barnes MR, McDonald CJ. Representating and Querying Conceptual Graphs with Relational Database Management Systems is Possible. In: American Medical Informatics Association Annual Symposium; 2000; 2000.
20. Hausam RR, Lu B. Conceptual Inferencing for Real-time Clinical Decision Support Using Hierarchical Queries in a Relational Database Management System. In: American Medical Informatics Association Annual Symposium; 2002; San Antonio; 2002.
21. Ryan G, Bernard H. Data Management and Analysis Methods. In: Denzin N, Lincoln Y, editors. The Handbook of Qualitative Research. Thousand Oaks: Sage Publications; 2000. p. 769.
22. Kayed A, Colomb RM. Extracting Ontological Concepts for Tendering Conceptual Structures. Data & Knowledge Engineering 2002;40:71-89.

CONCLUSION TO THE DISSERTATION

Heterogeneous databases represent a pervasive and persistent problem in a variety of domains, and, prior to this research, generalized, re-usable processes for their resolution were lacking. The difficulties in identifying such processes arise from a lack of characterizations for heterogeneities that can exist between multiple systems, and the lack of clarification of the goals and evaluation criteria for heterogeneous database integration. The first of the attached dissertation papers [4] provides both: a framework to identify representational heterogeneities, and a definitive goal for their resolution. The ultimate goal for heterogeneous database integration is to create a homogeneous data representation from heterogeneous source representations, while maintaining the data granularity and intended semantic meaning from each data source that are sufficient for the needs of the integrated data. Together, the framework for representational heterogeneities and the articulated goal of successful integration provide a means for identifying and evaluating current approaches, and organizing the field to solicit future research needs. This framework and articulated goal also supported the development of the generalized process for heterogeneous database integration that is the result of this research (Figure 1).

Figure 1. Process for Achieving Comparable Data from Heterogeneous Databases



This generalized process, presented in Figure 1 and discussed in the preceding papers [2, 5, 6], address the two very broad levels of heterogeneities – those resulting from different database schema, and those resulting from differences in the underlying data content – typically found across heterogeneous databases. [4] These two broad groupings of heterogeneities each became a target for integration processes, namely database schema integration and data content integration. Both require the use of a reference model. The information model (#2) is a referent model that assimilates different data structures (e.g., data instances, attributes, or tables) into a singular data element. Once the reference information model is selected or constructed (#2), local data structures are mapped, or transformed, to the new structural representation (#3). Heterogeneous data content (i.e., "what is in the fields") is made homogeneous by mapping the local data instances to concepts in a final reference terminology. The development of a reference terminology (#4), and associated mappings (#5 and #6) are addressed in [2]. The semantic focus of our process adds value to current syntactically-based efforts by suggesting a change in

focus from purely syntactical solutions toward a semantic-based approach, designed to capture the intended meaning and operational definitions of each data structure. Further, the generalized process addresses the importance of representing these differences in the final model to facilitate informed queries (#6) and analysis of the final data.

Together, the four preceding papers represent the spectrum of this dissertation research, from the problem definition, literature synthesis, and exploration of possible methodologies, to the actual development, implementation, and evaluation of the final generalizable process that was the proposed intent of the research. Different parts of this process are novel to different audiences. For example, the notion of a reference model, specifically a reference terminology, for content integration extends the current research boundaries of the database integration community, whose activities often stop at the stage of database schema integration. The guidelines for the development and thoughtful use of a reference terminology are also novel to the informatics literature, and the demonstration of the use of a reference terminology to increase the quality and expressiveness of aggregate health data is of tremendous value to the public health community who have strong interests in processes that facilitate the "re-use" of existing data sources for secondary analyses. The constructs and process introduced by this research as a whole are worthwhile to understand the complexity of issues and valid solution requirements for integrating heterogeneous data both in health care and in other domains.

The term comparability is used throughout the four preceding papers. Comparability is a broader notion than equivalence, and implies the need for a standard representation to make judgments of relationships between specific instance values (e.g., equivalent to; greater than, less than; broader than, narrower than, etc.) There are three general strategies to achieve comparability across multiple data sources: implicit, pair-wise, or reference model. Implicit strategies - the use of implied concept relationships that guide transformations of data to new representations - are the norm. But the conceptual frameworks underlying this strategy are buried within the psyche of the programmer or translator, and cannot be easily examined, validated, or changed. The second common approach is pair-wise – making comparisons or translations from one data representation

to another, exhausting all possible combinations in the database set. This strategy might ultimately provide more information, but is labor-intensive and less scalable, and cannot capture any similarities between the data sources nor facilitate comparisons between the source representations as a whole. The approach proposed here, that of an explicit outside reference model, allows a true assimilation of data from multiple representations to an explicit conceptual referent model that facilitates final information needs.

A major contribution of this research is the introduction and discussion of comparability as a goal and a requisite for heterogeneous database integration. The definition of comparability encompasses the examination of qualities and attributes that can facilitate determination of similarities and differences across multiple objects. A prerequisite for comparability is a homogeneous data representation. The thoughtful development of such representations, discussed in detail in [2] can satisfy final data needs and also capture the most (or most important) differences and similarities across local representations. As such, the use of the process created by this research can provide a true assimilation of local data, warts and all, to one representation. Theoretically, this representation can be used to create a representation for what is in use in multiple data collection systems, i.e., to describe "what is real". The application of this research can be used to survey and assimilate what data representations (both data structures and terms) are in use in a domain, and use the assimilation to develop or enhance data representation standards.

The reference model solution advocated throughout this work can be considered novel and perhaps radical to some, particularly in the informatics literature. The reference model for this demonstration consists of two parts: the Houston Asthma Reference Terminology (HART) that assimilates presenting complaint instances, and a global schema for emergency department (ED) visits that normalizes presenting complaint data structures into a single attribute and empowers the HART reference terminology with "context". Both parts of any standard reference model, if thoughtfully constructed, should capture similarities between the sources. In fact, one measure of an optimal reference model, and a next step for future research, is that which captures the most, or perhaps the most relevant, similarities across source representations. A strict data integration

approach, that simply combines every concept or combination of concepts from all local terms, as mentioned in [2], is often easy to construct but requires an implicit or informed understanding of the source representations to query and use in any meaningful way. In contrast, the focus of the process we have developed is on the capture of the intended semantics, or meanings, of the native data representations and the transformation of those semantics to a final homogeneous representation that is internally valid (e.g., non-redundant) and meets the needs of the final data.

Our strategy of using a reference model as a final homogeneous representation was particularly motivated by our primary objective to compare and combine data from multiple sources. Had our purpose been different, a pair-wise translation approach might have been sufficient. However, it is likely that in public health informatics, almost all projects will benefit from our reference model approach, since most public health activities require the aggregation of multiple data sources. To this end, our process can serve as a blueprint for achieving comparable and integrated data from many source representations and facilitate the reuse of existing data sources for secondary analyses. This approach has the advantage of dealing with reality – a variety of independently constructed and heterogeneous data representations - and capturing their similarities in a summary of sorts – an assimilation of concepts and semantic relationships that are captured in each of the component (native) representations. It is this assimilated reference model, or a bottom-up developed reference terminology, that could be used to extend existing terminology standards such as SNOMED to meet the real-world needs of public health and health services researchers.

The need to integrate and aggregate data from multiple sources is important to many health care (and non-health care) applications, and the use of a reference model can enable data integration in ways that retain local semantics and granularity, and capture similarities that exist between local representations. Once a homogeneous data representation is achieved, the data from heterogeneous databases can be compiled, shared, manipulated, and leveraged to address a multitude of information requirements.

This process recognizes that the differences inherent in native data representation and data collection contexts imply some loss of meaning or precision when being transformed to standard homogeneous representations. Key steps of this process attempt to minimize the loss of data semantics and granularity, potentially allowing better "quality" data in the final representation. This generalized process [5] should be valuable to a number of database integration efforts in a number of domains.

The specific informatics contributions of this research are: a framework for representational heterogeneities common in heterogeneous database integration projects, and a process that can be a blueprint for assimilating multiple, existing, heterogeneous data content for secondary analyses. Most notably, this process includes development guidelines for a reference terminology, a global schema or information model which gives context to a reference terminology, and a conceptual model and supporting scales for representing the quality of matching of heterogeneous local codes to a reference terminology. Future application of these generalized processes for the development and use of a reference terminology that were generated from this research increases the likelihood of informed use of the data and should be valuable in a variety of informatics applications.

From the start, the goal was to create re-usable knowledge, some artifact that showed promise to address other data integration problems, not just this one. The stated evaluation measures in the initial research proposal were simply to determine whether a process for homogeneous representation had been achieved and if the demonstration implementation succeeded. Additional evaluation criteria actually applied to this project include a critique of the final generalized process to determine if it addresses all of the representational heterogeneities presented in the first paper. Future potential evaluation targets for our generalized process include the validation of actual products (i.e., global schema and reference terminology reference models) and the integrated or comparable data that they enable. [6] This validation should describe the validity and the value of the transformed comparable data to address real information needs.

Future research is needed to determine if the process we have created can be repeated, and is both sound and useful. The combination of our heterogeneous data integration process, if validated in other projects, with explicit representations of context, should have potential for re-use and partial automation. Further research on the evaluation of the transformed and homogeneous data representations that result from the application of this process, both in other health care applications, and in other domains, is warranted and welcome.

The need to integrate data from multiple, heterogeneous source representations in health care is pressing and growing. The size and complexity of health care delivery and research activities, coupled with the lack of a-priori data representation and storage standards, has created a world of isolated data “silos” that to date cannot be analyzed in aggregate. Currently, the health care domain is overwhelmed with data that is largely incomparable, yet the needs for examining these data are becoming more urgent. Some of the rising costs of health care delivery and experimental drug development could be curtailed by using existing data sources and observational research designs on large populations. Similarly, evidence-based care, which requires monitoring data from multiple sources for long periods of time, could move from vision to reality if comparable data could be obtained across multiple populations and multiple points in the health care system. Issues of patient safety and health care quality are receiving well-deserved attention and driving needs to look at aggregate data from multiple sources to monitor health care activities and outcomes. Finally, new attention on bioterrorism surveillance and detection has drawn the spotlight on lack of integration of health care data for public health monitoring. The use of this generalized process to achieve comparable data has enormous potential to positively impact a plethora of health care quality and public health activities across the nation, and this research presents a starting point for aggregating data to support research and practice in health care and other domains.

REFERENCES

1. Sheth AP, Larson JA. Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. ACM Computing Surveys 1990;22(3):183-236.
2. Richesson RL, Turley JP, Johnson-Throop KA, Eick C, Tuttle MS. Development and Use of a Reference Terminology to Maintain Data Granularity and Semantics in the Integration of Heterogeneous Data Content. Submitted to: Data and Knowledge Engineering, 8-03 2003.
3. Campbell KE, Oliver DE, Spackman KA, Shortliffe EH. Representing Thoughts, Words, and Things in the UMLS. Journal of the American Medical Informatics Association 1998;5:421-431.
4. Richesson RL, Turley JP, Johnson-Throop KA, Eick C, Tuttle MS. Foundations for Heterogeneous Database Integration: A Framework to Identify Representational Heterogeneities. Submitted to: Journal of the Association for Computing Machinery, 8-03 2003.
5. Richesson RL, Turley JP, Johnson-Throop KA, Eick C, Tuttle MS. Creating Homogeneous Data from Heterogeneous Representations: A Process for Heterogeneous Database Integration. Submitted to: Data and Knowledge Engineering, 8-03 2003.
6. Richesson RL, Turley JP, Johnson-Throop KA, Eick C, Sockrider M, Macias CG, et al. Obtaining Comparable Presenting Complaint Data From Heterogeneous Emergency Department Databases. Submitted to: Journal of the American Medical Informatics Association, 8-03 2003.
7. Burgun A, Botti G, Fieschi M, Le Beux P. Issues in the Design of Medical Ontologies Used for Knowedge Sharing. Journal of Medical Systems 2001;25(2):95-108.
8. Sugumaran V, Storey VC. Ontologies for Conceptual Modeling: Their Creation, Use, and Management. Data & Knowledge Engineering 2002;42:251-271.
9. McGuinness DL. Conceptual Modeling for Distributed Ontology Environments. Proceedings of the Eighth International Conference on Conceptual Structures Logical, Linguistic, and Computational Issues (ICCS 2000) 2000(August 14-18, 2000).

Curriculum Vita

Rachel L. Richesson, PhD, MPH

ADDRESS:	University of Texas Health Science Center School of Health Information Sciences 7000 Fannin, Suite 600 Houston, TX 77030	14203 Cypress Meadow Dr. Cypress, TX 77429
PHONE:	713-500-3456	281-370-5685
E-MAIL:	Rachel.L.Richesson@uth.tmc.edu	

ACADEMIC PREPARATION:

2001-2003	University of Texas Health Science Center School of Health Information Sciences Houston, TX	Ph.D. Health Informatics

Concentrations: Public Health Informatics, Knowledge Representation, Health Care Terminologies, Heterogeneous Database Integration, Data Modeling
Dissertation: Knowledge Integration from Heterogeneous Health Care Data Sources

1998-2000	University of Texas Health Science Center School of Health Information Sciences Houston, TX	M.S. Health Informatics
1992-1995	University of Texas Health Science Center School of Public Health Houston, TX	M.P.H. Community Health Practice
1987-1991	University of Massachusetts Amherst, MA	B.S. Biology

PROFESSIONAL POSITIONS:

Mar. 2002-Present	National Library of Medicine Fellow University of Texas School of Health Information Sciences Houston, TX	Applied Health Informatics

1997 - 2001 Research Associate / Project Manager

Department of Behavioral Sciences
University of Texas School of Public Health
Houston, TX

1996 - 1996 Senior Research Assistant / Cohort Coordinator

Center for Health Promotion, Research, and Development
University of Texas School of Public Health
Houston, TX

1993 - 1995 Research Assistant / Information Specialist

AIDS Regional Education and Training Centers for Texas and Oklahoma
University of Texas School of Public Health
Houston, TX

PUBLICATIONS:

Articles:

Richesson RL, Turley JP, Johnson-Throop KA, Eick C, Tuttle MS. (2003) Foundations for Heterogeneous Database Integration: A Framework to Identify Representational Heterogeneities. *Under Review*, Journal of the Association for Computing Machinery, August 2003.

Richesson RL, Turley JP, Johnson-Throop KA, Eick C, Tuttle MS. (2003) Development and Use of a Reference Terminology to Maintain Data Granularity and Semantics in the Integration of Heterogeneous Data Content. *Under Review*, Data and Knowledge Engineering, August 2003.

Richesson RL, Turley JP, Johnson-Throop KA, Eick C, Tuttle MS. (2003) Creating Homogeneous Data from Heterogeneous Representations: A Process for Heterogeneous Database Integration. *Under Review*, Data and Knowledge Engineering, August 2003.

Richesson RL, Turley JP, Johnson-Throop KA, Eick C, Tuttle MS. (2003) Obtaining Comparable Presenting Complaint Data From Heterogeneous Emergency Department Databases. *Under Review*, Journal of the American Medical Informatics Association, August 2003.

Richesson RL, Turley JP. (2003) Triangulation Methods for Understanding the Construction of Conceptual Models**.** *Submitted for Publication*, Journal of Nursing Scholarship.

Richesson RL, Turley JP. (2003) Conceptual Models: Definitions, Construction, and Applications in Public Health Surveillance. Journal of Urban Health 80:suppl.

Richesson RL, Hwang L. (1998) Impact of the 1993 CDC Surveillance Definition of

AIDS in Texas, 1991-1994. Texas Medicine 94(1):56-63.

Richesson RL. (2000) Outcomes Research: A Review and Case for Outcomes Research Training for Health Informatics Professionals. Masters State of the Science Paper in Health Informatics, University of Texas School of Health Information Sciences, Houston, TX.

Markham C, Baumler ER, **Richesson RL**, et al. (2000) Impact of HIV-Positive Speakers in a Multi-Component, School-Based HIV/STD and Pregnancy Prevention Program for Inner-City Adolescents. AIDS Education and Prevention 12(4), May/June 2000.

Papers and Abstracts Presented at Conferences:

Mirhaji P, Turley JP, **Richesson RL**, Zhang J, Smith JW, Srinivasan A. Public Health Situation Awareness: A Semantic Approach. *Accepted for Presentation*, Conference on Multisensor, Multisource Information Fusion: Architectures, Algorithms, and Applications VIII, part of the International Symposium on Defense and Security, Orlando, FL, April 12-16, 2004.

Turley JP, **Richesson RL**, Johnson-Throop KA, Eick C, Tuttle MS. The Role of Context in the Integration of Heterogeneous Health Care Databases. *Submitted for Presentation*, MedInfo, 11th World Congress on Medical Informatics, San Francisco, CA, November 2004.

Richesson RL, Turley JP, Johnson-Throop KA, Eick C, Sockrider MM, Macias CG, Tuttle MS. Achieving Comparable Presenting Complaint Data from Heterogeneous Emergency Department Data. *Submitted for Presentation*, MedInfo, 11th World Congress on Medical Informatics, San Francisco, CA, September 7-11, 2004.

Richesson RL, Turley JP, Mirhaji P. The Use of Ontologies and Semantic Linkages in Public Health Surveillance. *Submitted for Poster Presentation*, MedInfo, 11th World Congress on Medical Informatics, San Francisco, CA, September 7-11, 2004.

Shankar P, Walji M, Kaur I, **Richesson RL**, Turley JP, Johnson-Throop K. Knowledge Modeling for Early Detection of Bioterrorism. *Submitted for Presentation*, MedInfo, 11th World Congress on Medical Informatics, San Francisco, CA, September 7-11, 2004.

Mirhaji P, Turley JP, **Richesson RL**, Zhang J. A Knowledge-Driven Approach to Public Health Situation Awareness. National Conference for Syndromic Surveillance, NYC, October 20-24, 2003.

Richesson RL, Turley JP, Johnson-Throop KA, Sockrider M, Macias CG, Tuttle MS. Creating Comparable Health Care Data for Public Health Surveillance and Analyses. *Accepted for Presentation*, American Public Health Association Annual Meeting and Exposition, San Francisco, CA, November 16-19, 2003.

Eriksen LR, Turley JP, Richesson RL. Handheld Device Improves Data Collection Accuracy. 8th International Congress in Nursing Informatics, Rio de Janeiro, Brazil, June 20-25, 2003.

Richesson RL, Turley JP, Tuttle MS. The Semantic Web and the Integration of Health Data Resources. American Medical Informatics Association Annual Symposium, San Antonio, TX, November 9-13, 2002.

Richesson RL, Turley JP, Riggs J, Miller J, Johnson C. The A-Z for Personal Digital Assistant (PDA) Applications in Health Care. American Medical Informatics Association Annual Symposium, San Antonio, TX, November 9-13, 2002.

Richesson RL, Turley JP. Conceptual Models: Definitions, Construction, and Applications in Public Health Surveillance. National Conference for Syndromic Surveillance, NYC, September 23-24, 2002.

Richesson RL, Wang W. The Use of Bayesian Networks and Decision Analysis in Solving Clinical Decisions for the Administration of Antiretroviral Agents to Prevent HIV Seroconversion Following an Occupational Percutaneous Exposure. American Medical Informatics Association Annual Symposium, Washington, D.C. November 7-9, 1999.

Richesson RL, Roberts RE, Roberts CR, Tortolero S, Click L. A Protocol to Identify and Manage Adolescents at Risk for Suicide and Abuse within the Context of an Epidemiologic Field Study. University of Texas - Houston Health Science Center Faculty Research Symposium, Houston, TX. October 8, 1999.

Riggs JW, **Richesson RL**, Niu K, Johnson TR. GynERConsult: Gynecological Emergency Room Data Management and Learning Tool. University of Texas - Houston Health Science Center Faculty Research Symposium, Houston, TX. October 8, 1999.

GRANTS/FELLOWSHIPS RECEIVED:

2001 – 2003 PI: National Library of Medicine F38 Applied Informatics Fellowship
"Knowledge Mapping Across Disparate Patient Care Datasets"
$129,415

2000 Pre-doctoral Research Fellow, Keck Center for Computational Biology.

PROFESSIONAL MEMBERSHIPS:

American Medical Informatics Association (AMIA)
American Public Health Association (APHA)
Institute of Electrical and Electronics Engineers (IEEE)
Texas Economic and Demographic Association (TEDA)

HONORS AND AWARDS:

2002 — 1st Place Winner, Student Poster Contest
University of Texas School of Health Information Sciences

1999 — Dean's Scholarship Award
University of Texas School of Health Information Sciences

COMMUNITY SERVICE:

2003 — Volunteer, Houston Trauma LINK Coalition

2003 — Participant, Developing a National Agenda for NHII, National Health Information Infrastructure 2003, Washington, D.C. June 30-July 2, 2003.

2003 — Reviewer, American Medical Association Annual Symposium, 2004.

2002 — Reviewer, Session on Biomedical Ontologies at the Pacific Symposium on Biocomputing 2003.

2001 – 2003 — President, UTSHIS Alumni Association

JOB-RELATED SKILLS:

Conceptual data modeling, research design and implementation, written and oral communication, knowledge of health informatics theory and practice, knowledge of public health theory and practice, database theory, database management, project management, community needs assessment, program evaluation, Semantic Web applications

Software: Java, Visual Basic, Fortran, SPSS, MS Access, Dbase, XML, RDF/OWL/Semantic Web tools, MS Project