Communications of the Association for Information Systems

Volume 17 Paper

May 2006

Advances in Data Modeling Research

Gove Allen
University of Minnesota, gallen@csom.umn.edu

Akhilesh Bajaj
University of Tulsa, akhilesh-bajaj@utulsa.edu

Vijay Khatri
Indiana University

Sudha Ram
University of Arizona

Keng Siau
University of Nebraska-Lincoln, siauk@mst.edu

Communications of the Association for Information Systems

ADVANCES IN DATA MODELING RESEARCH

Gove Allen
Tulane University

Akhilesh Bajaj
The University of Tulsa
akhilesh-bajaj@utulsa.edu

Vijay Khatri
Indiana University

Sudha Ram
University of Arizona

Keng Siau
University of Nebraska-Lincoln

ABSTRACT

In this paper, we summarize the discussions of the panel on “Advances in Data Modeling Research,” held at the Americas Conference on Information Systems (AMCIS) in 2005. We focus on four primary areas where data modeling research offers rich opportunities: spatio-temporal semantics, genome research, ontological analysis and empirical evaluation of existing models. We highlight past work in each area and also discuss open questions, with a view to promoting future research in the overall data modeling area.

I. INTRODUCTION

Commercial database management systems (DBMS) have been available since the 1960s. Earlier models included the network and the hierarchical models (Bachman 1969; Tsichritzis et al. 1976). In the 1970s, the relational model contributed to the separation of the logical representation of data (relations and tuples) from the physical implementation (files and access mechanisms) (Codd 1970). One of the shortcomings of the relational model was its limited ability to explicitly capture requirement semantics (Hull et al. 1987). Several semantic data models have been proposed in order to overcome this limitation. Examples include the Entity Relationship model (ERM) (Chen 1976), extended ERM (Smith et al. 1977), the NIAM model (Nijssen et al. 1989; Verhijen et al. 1982), TAXIS (Mylopoulos et al. 1980), Semantic Data Model (SDM) (Hammer et al. 1981), and the Unified Semantic Model (USM) (Ram 1995). Semantic data models generally consist of richer concepts such as generalization, aggregation, classification, relationship representation, and in some cases, event modeling, in order to explicitly map to semantics in the requirements. While systems that support semantic model implementation have largely consisted of research prototypes, in many cases, semantic models have been utilized as initial models in the systems analysis phase, for capturing user requirements, prior to creating a relational schema (Korth et al. 2005).

Over the last decade, data modeling research has diversified into several subareas. First, several researchers have proposed new models or extensions to existing models, such as the relational model, that explicitly model temporal and spatial semantics in the requirements domain (spatio-temporal data models). The goals here have ranged from providing assistance in modeling at the systems analysis level, all the way down to the more efficient implementations at the physical level.

Second, several domains have arisen that require specialized modeling and retrieval. For example, in the area of genomic research, the semantics of the underlying data, as well as the processes used by researchers who utilize this data, can benefit from novel concepts at the data model level. Concepts that capture the underlying relationships between genes in a sequence, as well as the creation of query operators to specifically model the kind of queries bio-researchers most often formulate, can potentially improve both the underlying representation of genomic data as well as assist in the formulation and execution of queries.

Third, there has been increasing incorporation of ontologies in the area of data model evaluation. Most commonly, the Bunge ontology (Wand et al. 1995) has been used in providing a set of philosophically robust concepts that were formulated for generalized descriptions of our world. These concepts have been utilized to evaluate the completeness, ambiguity and correctness of existing data models.

Finally, there has been renewed interest in the empirical evaluation of data models in the area of systems analysis. Questions in this area include the development and testing of theories that potentially apply to systems analysts when they model requirements using data models, as well as prescriptions of new concepts that can enhance existing data models from a systems analysis perspective.

The primary purpose of this work is to summarize the discussions of the panel on “Advances in Data Modeling Research” at the 2005 *America’s Conference on Information Systems (AMCIS).* The panel presented illustrative work in all four of these emerging areas and highlighted the research questions that remain unanswered.

The rest of this paper is organized as follows. In Section 2, we discuss emerging work in the area of spatio-temporal data models. Section 3 describes the semantics that occur in genomic research, and provides examples of how semantic model concepts and query operators can be used to increase the efficiency of data modeling and retrieval for this important domain. Section 4 describes recent and emerging work in the application of the Bunge ontology to data models. Section 5 describes opportunities for empirical work in the evaluation of data modeling. We conclude with a discussion and opportunities for future research in Section 6.

II. SPATIO-TEMPORAL DATA MODELS

Most applications require some aspect of time in organizing their information, for example, healthcare (patient histories), insurance (claims and accident histories), reservation systems, and scientific applications. Geospatial information has been applied to business applications such as facility management, market analysis, transportation, logistics, strategic planning, and decision-making (Mennecke et al. 1996). Underlying these applications are temporal and geospatial data, collectively referred to as *geospatio-temporal* data.

In developing geospatio-temporal applications, there is a need to elicit the data semantics not only related to “what” is important for the application but also related to “when” and “where.” One of the problems with designing such applications is that there is “a gulf between the richness of knowledge structures in application domains and the relative simplicity of the data model in which the structures can be expressed" (Worboys et al. 1990). *Conventional conceptual models* (Chen 1976; Elmasri et al. 1994) that provide a formalism to represent "what" is pertinent for an application are only partially useful for geospatio-temporal applications. To help represent the geospatio-temporal data requirements, there is a need for geospatio-temporal conceptual models.

CHARACTERIZING GEOSPATIO-TEMPORAL DATA SEMANTICS

Prior research (Shoval et al. 1994) suggests that a conceptual model should be powerful in "semantic expressiveness," where expressiveness refers to the availability of a large variety of concepts for a more comprehensive representation of the real world (Batini et al. 1992). A geospatio-temporal conceptual model needs to elicit data semantics related to time and space like *event* and *state* (Jensen et al. 1998), *valid time* and *transaction time* (Snodgrass et al. 1986), *lifespan* (or *existence time*) (Gregersen et al. 1998), *temporal* and *geospatial granularities,* along with *indeterminacy* (Khatri et al. 2002), and *geometry* and *position* (David et al. 1996). We provide brief definitions of these terms below; for more details on the geospatio-temporal data semantics, the reader is referred elsewhere (Khatri et al. 2002; Khatri et al. 2004).

An event occurs at a point in time, that is, it has no, or an extremely short, duration (for example, lightning hit the road at 2:03 PM), while a state has duration (for example, a storm occurred from 5:07 PM to 5:46 PM). While valid time denotes when the fact is true in the real world and implies the storage of histories related to facts, transaction time links an object to the time it is current in the database and implies the storage of versions of a database object. Existence time, which applies to an object, is the valid time when the object exists. Position in space is based on coordinates in a mathematically-defined reference system, for example, latitude and longitude. The shape of the object is represented by geometry, for example, point, line, and region. Granularities are intrinsic to geospatial and temporal data, and provide a mechanism to hide details that are not known or not pertinent for an application. For example, in a cadastral application (Hermosilla 1994), mortgages can be associated with a temporal granularity of `day`, and the representation of long-term land-use changes may require a temporal granularity of `year`. `Day` and `year`, or more accurately `Gregorian day` and `Gregorian year`, are examples of temporal granularities, which belong to the Gregorian calendar.

Having briefly outlined the data semantics that need to be elicited using a geospatio-temporal conceptual model, we next describe the criteria for augmenting conventional conceptual models.

CRITERIA FOR AUGMENTING CONVENTIONAL CONCEPTUAL MODELS

Prior research (Böhlen et al. 1998) proposes requirements to ensure that legacy DBMS application code (along with the data) remain operational when migrated to geospatio-temporal DBMS. While the requirements—upward compatibility and snapshot reducibility—described in their paper refer to logical data model and query languages, the requirements are equally applicable to conceptual modeling.

Upward compatibility (Böhlen et al. 2000; Snodgrass et al. 1997) implies the ability to render conventional conceptual schemas geospatio-temporal without affecting the legacy schemas. The objective of upward compatibility is to be able to develop geospatio-temporal schemas without invalidating the extant legacy schemas, thus, helping protect investments in existing schemas. It also implies that both the legacy schemas and the geospatio-temporal schemas can co-exist. Upward compatibility requires that the syntax and semantics of the traditional conceptual model (*e.g.,* Chen 1976; Elmasri et al. 1994), remain unaltered. If the geospatio-temporal extension is a strict superset provided by adding non-mandatory semantics, it would ensure that the geospatio-temporal extension is upwardly compatible with conventional conceptual models.

Snapshot reducibility (Böhlen et al. 2000; Snodgrass 1987) refers to "natural" generalization of the semantics of extant conventional conceptual models, *e.g.,* for incorporating the geospatio-temporal extension. Snapshot reducibility ensures that the semantics of a geospatio-temporal model are understandable *in terms of* the semantics of the conventional conceptual model. Here, the overall objective is to help ensure minimum additional investment in training (or retraining) a data analyst.

We next describe a geospatio-temporal conceptual model that considers the requirements stated above.

REPRESENTING GEOSPATIO-TEMPORAL DATA SEMANTICS

Prior research (see, for example, Khatri et al. 2004; Snodgrass 1999) argues that, in order to simplify the complex task of representing geospatio-temporal data semantics, geospatio-temporal aspects should be the *last* consideration in conceptual design (see Figure 1). As shown in Figure 1, there are two levels of abstraction,[1] one for "what" and the other for "when/where." Current thinking suggests employing a supplementary level of abstraction for representing the geospatio-temporal data semantics; this supplementary level is typically provided via geospatio-temporal annotations (see, for example, Gregersen et al. 1998; Khatri et al. 2004; Parent et al. 1999; Tryfona et al. 1999).

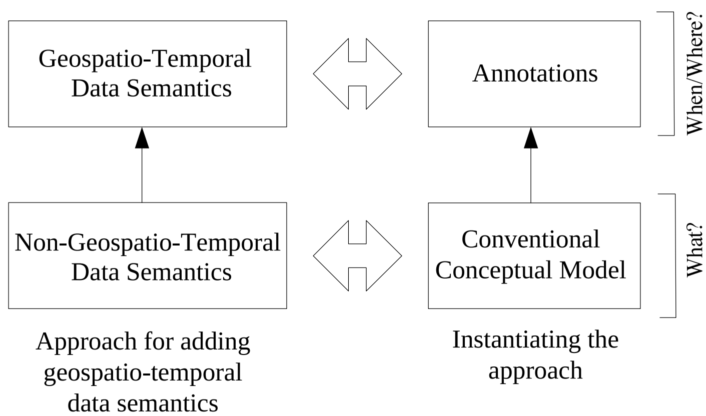

Figure 1: Representing Geospatio-temporal Data Semantics

To illustrate an annotation-based conceptual model that represents geospatio-temporal data semantics on the schema, we provide an example using ST (Spatio-Temporal) USM (Khatri et al. 2004) (see Figure 2). In a conventional conceptual model, a rectangle is used to represent *entity types* (Elmasri et al. 1994). Figure 2, for example, denotes that "`LAND PARCEL`" is an entity type that is pertinent to a database application. In this example, "`LAND PARCEL`" is represented as a region and has an associated lifespan (or existence time). ST USM employs a textual string to denote existence time, which is represented as state ("S"); in Figure 2, the time periods have a temporal granularity of "`day`." Additionally, "`LAND PARCEL`" is represented as a region ("R") in the horizontal plane with a geospatial granularity of "`deg`" (i.e., degree).

[1] An abstraction provides a mechanism for focusing on selected details while deliberately omitting others.

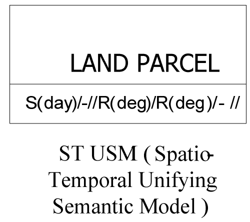

Figure 2: Capturing the Geospatio-Temporal Semantics using Annotations

Note that ST USM is based on USM (Ram 1995), an extended ER Model (Chen 1976). Similar to the entity type, other types of constructs (*e.g.,* attribute, subclass, aggregate class) can also be annotated. We present the explicit data semantics associated with these geospatio-temporal annotations elsewhere (Khatri et al. 2002; Khatri et al. 2004; Khatri et al. 2006).

Annotated schemas do not invalidate legacy conceptual schemas as the syntax and semantics of conventional conceptual models is unaltered, i.e., annotated schemas are upwardly compatible. Upward compatibility implies that annotating the schema would induce the sequenced geospatio-temporal semantics; on the other hand, removing the annotations would render the schema with the traditional (snapshot) semantics. For example, in a conventional conceptual model, a *key attribute* (Elmasri et al. 1994) uniquely identifies an entity (at a point in time). A *temporal key* (Snodgrass 1999) implies uniqueness *at each point in time*. As may be evident, the semantics of a temporal key here are implied by the semantics of a key in a conventional conceptual model; thus, an annotated schema is snapshot reducible. In summary, annotations provide a succinct way of representing the geospatio-temporal aspects that are important for temporal and geospatial applications.

Further work would be useful in several areas. It would be helpful to explore how ST USM can be used as a canonical model for information integration of distributed geospatio-temporal data. The annotations should be extended to incorporate schema versioning as well as to provide a mechanism for modeling geospatio-temporal constraints in a conceptual schema, such as lifetime constraints and topological constraints.

III. MODELING THE SEMANTICS OF BIOLOGICAL SEQUENCES

As genomic research is becoming mature, large amounts of sequence data—including DNA (Nucleotide) sequence and primary protein sequences—are being generated and stored in various sequence databases. A sequence by itself is not informative; it must be analyzed by comparing against existing databases to develop hypotheses concerning relationships and functions. Currently, a substantial portion of a biological researcher's daily routine is spent in analyzing sequence data. For example: An abundant message in a cancer cell line may bear similarity to protein phosphatase genes. This relationship would prompt experimental scientists to investigate the role of phosphorylation and dephosphorylation in the regulation of cellular transformation. Meanwhile, the exponential growth in biological data in the past few decades has added a predictive element to research in this field. Biological data and events can be explained and analyzed using basic physical and chemical laws. These kinds of analyses will facilitate the discovery of new biological trends and laws crucial to our understanding of these complex systems.

Computational tools and databases are essential to the management and identification of subtle similarities and patterns found in the exponentially growing volume of biological sequence data. Two major components of this data are DNA sequences and protein primary structure data. Two of the most well-known and reliable biological sequence data repositories are: The National Center for Biotechnology Information (NCBI) in the United States (Woodsmall et al. 1993) and the European Bioinformatics Institute (EBI) (Emmert et al. 1994) in England. The sequence data stored at these two places adopt somewhat similar formats (Benson et al. 2000)—the sequence data itself is stored as string-like text attribute for each sequence entity together with other biological information concerning the sequence. The semantics of the sequence data is not explicitly expressed, and this prevents researchers from directly running queries to analyze the data. To search, compare, and analyze the sequences, a general approach involves the use of a set of algorithms such as BLAST (Basic Local Alignment Search Tool) (Altschul et al. 1990), which was developed in 1990. Multiple programs have been implemented using BLAST or modified BLAST algorithms (Altschul et al. 1997) which allow users to compare a DNA or protein sequence with all sequences in a specified database. Comparisons are made in a pair-wise fashion. Each comparison is given a score reflecting the degree of similarity between the sequences being compared. The higher the score, the greater is the degree of similarity. Many researchers have worked to improve the performance of sequence comparison. Much of this research has focused on algorithm development or application software development. Sequence comparison algorithm development has concentrated on improving the sensitivity of detecting remote sequence homologous relationships. Sequence comparison application software development has concentrated on improving the throughput of sequence comparison tasks and user exploration of sequence comparison results.

Despite all the progress from this stream of research, sequence search and comparisons are still very time-consuming and labor-intensive. Consider a user who wants to run the BLAST algorithm for a batch of sequences against a large set of target sequences. First, the researcher has to choose a program, usually a derivative of BLAST, and a database to search. Then the sequences for comparison have to be converted into the required format from a text file and loaded into the program. During the process of alignment, the BLAST program has to be executed iteratively for each comparison. The results are then combined into a report for the users. Finally, any valuable findings have to be loaded back to the databases for future reference. All of these steps are very time consuming, and one of the biggest challenges with current systems is the difficulty in performing analyses and queries directly on the stored sequence data. The most important reason for this challenge is that the semantics of the sequence data are not captured explicitly in the database. Semantics of sequential data have been researched in other fields or domains (Curtis et al. 1992; Segev et al. 1987; Seshadri et al. 1995), including business processes and temporal databases, in which the data appear to be similar to biological sequence data. Several data models and query languages have been developed to deal with these types of data (Seshadri et al. 1995). However, they are not adequate to represent the semantics of biological sequences nor are their operators sufficient to manipulate biological sequences. We believe it is important to focus on developing operators to query biological data based on a model that can explicitly represent the semantics of biological sequences including DNA and protein sequences. We believe that by explicitly expressing the meaning of the data stored in biological databases, it will become easier to develop new operators to query this data directly in ways that support the needs of biosequence researchers (Ram et al. 2003; Ram et al. 2004). Such semantics should explicitly represent the structure of sequences including the elements of the sequence, the position of each element, the relationships between various subsegments of each sequence and the biological functions of different segments of a sequence as well as of the complete sequence. Further, some types of sequences such as proteins are not merely linear—they fold into various kinds of three dimensional structures. The type of 3D folding then determines their biological properties and functions. We need new constructs to represent the semantics of these types of 3D sequences (Ram et al. 2004). Such semantics will allow us to define new operators to manipulate biological sequence data. These operators will allow users to ask and get answers easily for meaningful questions. The operators can be implemented directly in a relational DBMS so that users are able to query the data more easily, resulting in a significant savings of end user time and other resources.

IV. ONTOLOGICAL RESEARCH IN DATA MODELING

One recent advance in data modeling research is the application of Ontology (the branch of philosophy that deals with the description of existence). By far, the philosopher most cited in such works in information system literature is Bunge (1977). Bunge's work is very attractive to information systems researchers because it is, unlike the ontological works of many other philosophers, built on a firmly scientific foundation. Moreover, elements of Bunge's work correspond very nicely to basic constructs used in conceptual modeling for decades.

Bunge's intent was to develop a formalization that could be used to describe the existence of the universe and all components that comprise it. As such, he focuses on things and their states as the building blocks for his ontology. Because the processes that cause change in natural systems are often ill-understood, his approach is particularly well-suited to natural systems because it focuses on states and changes of states as they are observed. Accordingly, his ontology lacks constructs to represent the rules that govern how state changes from one point in time to another in response to some occurrence happening in the world. In fact, it lacks constructs to represent the occurrences themselves. This does not mean that Bunge's ontology is deficient in any regard, only that, like many ontologies, its scope is limited to the description of the current and past states of things.

Such a conceptualization works well for the description of natural systems. Natural systems are comprised of material (physical) things. Material things have properties that can, in principle, be observed. This category of things need not exist in nature per se, that is, it also includes things created by human intention—as long as such artifacts are material. The defining characteristic of natural systems is that things change states in accordance with natural laws. It is the goal of many areas of natural science to understand the natural laws that govern the state changes of things in natural systems. For example, early in the study of classical mechanics, Newton proposed three laws of motion (dealing with inertia, momentum, and action/reaction) and proved that these laws governed the motion of both everyday and celestial objects. These laws predict how the state of a thing (i.e. its velocity) will change when known forces are applied to it.

Artificial systems are fundamentally different from natural systems because they are largely concerned with attributes that are ascribed to things rather than with properties, which are inherent to things. To illustrate the difference, consider the height and the name of a person. Height is a property. It is intrinsic to the individual. Although it can be expressed in different scales, it can be measured. Conversely, name is an attribute—it is a value that is "attributed" to a person. It cannot be measured because it is not intrinsic to the individual. The value of the name attribute for an individual is generally set shortly after birth, and may, like height, change during the life of the individual. However, the value of the individual's height property changes according to natural laws, while the value of the name attribute changes according to socially constructed rules. Often the value of the name attribute changes when an individual marries, but the rules that govern how the value will change are artificial. Artificial rules are constructed by human intentionality, vary from society to society, and are categorically defined in relationship to some set of causal occurrences or events.

To a large degree, business systems are artificial systems. They are not concerned with measurement of intrinsic properties; rather, they seek to record how values in the domain of discourse change as a result of occurrences that pertain to the business. For example, a business information system may need to report the value of the "revenue" attribute for a company during a given period. Revenue cannot be directly observed or measured because, like the "name" attribute, its value is defined by accepted societal rules that map a set of causal events to the attribute's value. In this case, the events that determine the attribute's value are "sale" events or "receive payment" events depending on whether the revenue attribute is to be calculated on the cash basis or on the accrual basis. Of course, the revenue attribute is not alone. Virtually all of the values reported on standard financial statements are attributes and cannot be observed. Rather, they must be inferred by applying a set of rules to the set of relevant occurrences that have take place.

The very reason that information systems record transactions is to store information about the events that must be summarized to infer the values of attributes that cannot be measured, or are difficult to measure. Clearly, the financial position of a business is not the only example. All systems that track ownership of assets must also record how events affect the socially constructed attribute of ownership. Judicial systems need to record various events about individuals to be able to apply rules for the sentencing of convicts. Even when asserting state is not the primary function of an information system, there may be valid reasons to conceptualize it as an event tracking mechanism. In healthcare systems, quite apart from the need to generate bills, physicians need access to treatment histories to be able to make informed diagnoses and prognoses as well as to recommend effective treatment.

The role of ontology in data modeling is to guide the modeling process with a particularly philosophy of how to describe things and their interactions. For example, Bunge's ontology asserts that things either have an attribute with a value or they do not have that attribute. That is, the ontology does not allow a thing to have an attribute with a null value (Gemino et al. 2005). When constructing a data model informed by this ontology, a modeler would choose to represent similar things with different attributes in a class hierarchy rather than in a single class with attributes which may or may not have a value for individual members of the class.

When an ontology lacks the expressiveness to describe the rules that govern socially constructed attributes, it must treat attributes and properties uniformly. This means that the same state-tracking approach that is appropriate for recording the values of properties without reference to the natural laws that govern their changes must also be used for recording the values of attributes—even though rules that determine the value of the attributes are central to the system that must be developed. The result is to design the data structure that will support the information system without referencing the events that are essential to the application of the rules that ascribe values to attributes. This does not mean that ultimately the system will not be able to produce the values of the required attributes, only that the events that cause the attributes to take on certain values will not be recorded as a part of the information system's data.

When the only goal of an information system is to report on the most current state (the most recent value of each attribute in the system), this distinction is of little importance; however, when the system also needs to report states for arbitrary points in history, the distinction is monumental. When a system records the most recent state for a property, the property is measured and the value is recorded. When an attribute is treated like a property, an event is observed, a rule is applied to determine the new value, which is then recorded. In this sense, things have attributes and properties, which are treated uniformly, and only their most recent value is accessible. However, when past states must be accessible in addition to current state, the distinction becomes clear. For a property, the system has only had access to its vector of values over time, so there is no option other than to record that set of values as a state history. Clearly, for attributes calculated by the system, the system has access to the same vector and can also record a simple state history; however, the system also has access to the events that cause state to change as well as to the rule that governs the state change. To record only the state history is to lose the causal record of why the attribute achieved different values at different times. With a property, cause is unknown, so recording only the state history does not result in data loss; however, with an attribute that has its value determined by the system, not only is causality known, it is already implemented somewhere in system logic. Accordingly, when its value over time is recorded as a state history, information about causation is needlessly discarded.

Consider a system that maintains the current credit limit for customers. Over time, various events occur that cause the limit for a given customer to change. The events are governed by rules. For example, a rule might state that if a customer requests a credit limit and her limit has not changed for 12 months and her current balance is currently zero, a five percent increase in credit limit occurs. Another rule might indicate that when a sales person requests a credit limit increase on behalf of a customer, any amount between zero and ten percent can be processed as long as no change has happened for the prior six months. If a five percent increase in a customer's credit limit is recorded, and that customer had no change in credit limit for twelve months or more, recording only the historical values of the attribute leaves no way of knowing whether the cause of the increase was a customer initiated request or a salesperson initiated request.

Whether the data model directly represents the events that cause the state of attributes to change, or merely represents changes in attribute values using a state-history structure, depends largely on how the modeler conceptualizes the interaction among things in the world. The role of ontology in information systems is to formalize rules about how the interaction of things in the world should be conceptualized. As such, the application of ontology can have a significant influence on how information systems are ultimately implemented.

Largely, research in data modeling regarding ontology can be classified into three broad categories. These are:

- research that describes an ontology-enlightened data modeling approach (*e.g.*, Wand et al. 1995),
- research that evaluates different aspects of traditional data modeling in light of ontological perspectives (*e.g.,* Burton-Jones et al. 1999; Wand et al. 1993), and
- empirical evaluations of user performance with artifacts developed under different ontological assumptions (*e.g.,* Allen et al. 2006; Bodart et al. 2001).

What our field is missing is the development of an ontology that is uniquely fitted to the demands of artificial systems. For research in ontology to meet its potential in informing the task of data modeling, we must develop, perhaps in cooperation with the field of philosophy, an ontology that addresses the unique problems posed by the conceptualization of artificial systems.

V. EMPIRICAL WORK IN DATA MODEL EVALUATION

A survey of the literature on the evaluation of modeling methods reveals several desirable attributes for conceptual modeling methods, which have been used as dependent variables in past empirical studies. These include

a) the adequacy or *completeness* of the modeling method in being able to represent the underlying reality (Amberg 1996; Bajaj et al. 1996; Brosey et al. 1978; Kramer et al. 1991; Mantha 1987; Moynihan 1996; Siau 2004),

b) the *readability* of the modeling method’s schemas (Hardgrave et al. 1995; Shoval et al. 1987; Siau et al. 1997), and

c) how easy it is to use the modeling method to *represent requirements* (Bock et al. 1993; Kim et al. 1995; Kramer et al. 1991; Shoval et al. 1987; Siau et al. 2005).

Batra et al. (1992) present an excellent summary of the early work in the area. More recently, Wand et al. (2002) and Gemino et al. (2001) have highlighted the several dimensions along which empirical work can be pursued in the area, while Topi et al. (2002) present a summary of recent empirical studies.

As an example dependent variable, the **readability** of a modeling method essentially indicates how easy it is to read a model schema and **reconstruct** the underlying domain reality from the schema. Readability is desirable in situations where the model schemas are created by one team of analysts and then need to be read and interpreted by other analysts, system developers or maintenance administrators during the course of the system’s lifecycle. For example, if a new database administrator requires an understanding of the schemas of existing database applications in the organization, then the readability of the model schemas that were created during the earlier analysis phases of the projects becomes important. Next, we examine the independent variables that have been considered in earlier work.

The first independent variable is the level of experience and familiarity of the subjects with the conceptual model used. Readers who are more experienced in the underlying conceptual model are thought to perform better at interpreting the schemas as well. In most studies (Brosey et al. 1978; Hardgrave et al. 1995; Palvia et al. 1992; Peleg et al. 2000), this variable has been controlled, by using subjects with similar backgrounds for all treatment levels. Second, past studies have attempted to control for the level of familiarity with the domain by utilizing domains that are reasonably familiar to all subjects, and further by randomly allocating subjects across treatment levels. A random allocation reduces the likelihood of small differences in domain familiarity between subjects in different treatment levels. A third variable is the underlying complexity of the requirements for a particular situation, where a more complex set of requirements is harder to reconstruct than a simpler set. This is controlled by utilizing the same requirements case across treatments (Juhn et al. 1985; Kim et al. 1995; Peleg et al. 2000).

A fourth independent variable whose effect has been studied is the modeling method itself, with the variables discussed earlier being controlled. While the results of earlier empirical studies have shown if one model's schema is more readable or more conducive to schema creations than that of another model, there has been very little attempt to explain **why** any differences were observed. There has been a lack of a theoretical basis for the hypotheses that were examined in empirical work, and for explaining findings. For example, finding that the extended ERM (EER) schema is more or less readable than the object–oriented (OO) model (Booch 1994) schema for a case does not indicate why this was observed. The problem is that existing models view reality in differing ways and hence differ from each other along several dimensions. Hence, it is difficult to isolate what aspect of a model may cause more or less readability.

One possible solution is to identify a set of universal attributes of all models, and then consider treatments that differ along one of these universal attributes. One step in this direction is the ontological framework called the Bunge Wand Weber framework (BWW) (Wand et al. 1995). The BWW framework utilizes an underlying ontology for all information systems. It then compares existing information system models on the basis of the degree to which concepts (or constructs) in the model and the ontology match. For example, a model that does not contain sufficient concepts to capture all of the underlying reality is termed to have *construct deficit*. Another approach is to map IS models to a meta-model. For example, the ER, OR (object relational) and UML (unified modeling language) models are compared in Halpin (2004).

A complementary approach for identifying a set of universal attributes is to consider measurable properties that can be applied to all models. The most obvious example of this kind of universal attribute is the number of concepts in a model: a property which is common to all models and easily measured (Bajaj 2004).

Finally, an alternative approach is to use cognitive theories (Bajaj et al. 2004; Siau 1999) to provide a guidance for empirical work in the area. Thus, cognitive theories such as the theory of short-term memory, pattern recognition and learning theory can be used to formulate hypotheses that guide the conduct of empirical studies and to explain the results.

VI. DISCUSSION AND CONCLUSION

In this paper, we have summarized the discussions of the AMCIS 2005 panel on advances in data modeling, sponsored by the Special Interest Group on Systems Analysis and Design (SIGSAND). While data modeling is a very well established area of research, the panel's discussions illustrated that there are still many rich opportunities for research in data modeling.

New areas that require data modeling research are continuously emerging and include spatio-temporal modeling, genomic research and agile modeling.

In the area of spatio-temporal representation, the issues include how to represent time and space abstractions in the data model in a generalizable sense that is applicable across application domains. A guiding constraint of the research in this area is upward compatibility of the legacy conceptual domain schema with any new spatio-temporal schema that may replace it.

In biological sequence modeling, one of the key challenges highlighted at the panel was the development of a framework to a) capture the semantics of genomic sequence data, and b) query datasets of genome sequences with special query-language operators that build on these semantics.

In the ontology area, one key challenge remains the extension of existing ontologies, that deal largely with inherent attributes in natural systems, so that these ontologies can explicitly represent non-intrinsic attributes and the rules that can change them. This will allow the onotology to depict not only the structure but also the process-logic of the artificial system.

Empirical work in data model evaluation has been largely observational in nature, and has only recently started trying to address *why* differences in modeling performance or comprehension are observed. The panel discussion highlighted possible approaches such as the BWW framework, cognitive theories and universal shared attributes that can be used to guide hypothesis formulation and explain findings from future empirical studies. Techniques such as cognitive mapping, action research and other non-laboratory methods can complement the controlled experiments that form the bulk of the existing work in the area.

We would like to emphasize that the list is by no means comprehensive and there are other emerging areas such as multimedia data and semi-structured data that provide data modeling opportunities as well. In each of the areas discussed by the panel, we attempted to provide a flavor for questions that remain unanswered. It is our hope that this work will serve as an impetus in the area of data modeling research in the IS community.

REFERENCES

Allen, G.N., and S.T. March (2006) "The Effect of State-based and Event-based Data Representations on User Performance in Query Formulation Tasks," *MIS Quarterly* (In Press).

Altschul, S.F., et al. (1990) "Basic Local Alignment Search Tool," *Journal of Molecular Biology* (215:3), pp 403-410.

Altschul, S.F., et al. (1997) "Gapped BLAST and PSI-BLAST," *Nucleic Acids Research* (25:17), pp 3389-3402.

Amberg, M.A. (1996) "A Pattern Oriented Approach to a Methodical Evaluation of Modeling Methods," *Australian Journal Of Information Systems* (4:1), pp 3-10.

Bachman, C.W. (1969) "Data Structure Diagrams," *Data Base* (1:2), pp 4-10

Bajaj, A. (2004) "The Effect Of the Number of Concepts On the Readability of Schemas: An Empirical Study With Data Models," *Requirements Engineering* (9:4), pp 261-270.

Bajaj, A., and S. Ram (1996) "A Content Specification for Business Process Models," *Australian Journal of Information Systems* (4:1), pp 22-31.

Bajaj, A., and S. Rockwell (2004) "COGEVAL: A Propositional Framework Based on Cognitive Theories to Evaluate Conceptual Models," Proceedings of the Workshop on the Evaluation of Modeling Methods in Systems Analysis and Design (EMMSAD).

Batini, C., S. Ceri and S.B. Navathe (1992) *Conceptual Database Design: An Entity-Relationship Approach* Benjamin/Cummings Publishing Company.

Batra, D., and J.G. Davis (1992) "Conceptual data modeling in database design: similarities and differences between novice and expert designers," *International Journal of Man-Machine Studies* (37), pp 83-101.

Benson, D., J. Lipman and J. Ostell (2000) "GenBank," *Nucleic Acids Research* (21:13), pp 2963-2965.

Bock, D., and T. Ryan (1993) "Accuracy in Modeling with Extended Entity Relationship and O-O Data Models," *Journal of Database Management* (4:4), pp 30-39.

Bodart, F., et al.(2001) "Should Optional Properties be Used in Conceptual Modelling?," *Information Systems Research* (12:4), pp 384-405.

Böhlen, M., C.S. Jensen and B. Skjellaug (1998) "Spatio-temporal database support for legacy applications," Proceedings of the 1998 ACM symposium on Applied Computing, Atlanta, Georgia, United States, pp. 226-234.

Böhlen, M.H., C.S. Jensen and R.T. Snodgrass (2000) "Temporal Statement Modifiers," *ACM Transactions on Database Systems* (25:4), pp 407-456.

Booch, G. (1994) *Object Oriented Analysis and Design with Applications* Redwood City:Benjamin/Cummings

Brosey, M., and B. Schneiderman (1978) "Two Experimental Comparisons of Relational and Hierarchical Database Models," *International Journal of Man Machine Studies* (10), pp 625-637.

Bunge, M.(1977) *Ontology I: The Furniture of the World* D. Boston, MA: Reidel.

Burton-Jones, A. and R. Weber (1999) "Understanding Relationships with Attributes in Entity-Relationship Diagrams," Proceedings of the 21st ICIS, Charlotte, NC, pp. 214-228.

Chen, P.P. (1976) "The Entity-Relationship Model - Toward a Unified View of Data," *ACM Transactions of Database Systems* (1:1), pp 9-36.

Codd, E.F. (1970) "A Relational Model for Large Shared Databanks," *Communications of the ACM* (13), pp 377-387.

Curtis, W., M.I. Kellner and J. Over (1992) "Process Modeling," *Communications of the ACM* (35), pp 75-90.

David, B., M.V.D. Herrewegen and F. Salge (1996) "Conceptual Models for Geometry and Quality of Geographic Information," in: P.A. Burrough and A. Frank (eds.) *Geographic Objects With Indeterminate Boundaries,* , Taylor & Francis, pp. 352.

Elmasri, R., and S.B. Navathe (1994) *Fundamentals of Database Systems*, (Second ed.) Redwood City, CA: Benjamin/ Cummings Publishing Co.

Emmert, B., et al.(1994) "The European Bioinformatics Institute (EBI) databases," *Nucleic Acids Research* (22:17), pp 3445–3449.

Gemino, A., and Y. Wand (2001). "Towards Common Dimensions In Empirical Comparisons of Conceptual Modeling Techniques," Seventh CAiSE/IFIP-WG8.1 International Workshop on the Evaluation of Modeling Methods in Systems Analysis and Design, Toronto, Canada, pp. 144-151.

Gemino, A., and Y. Wand (2005) "Complexity and Clarity in Conceptual Modeling: Comparison of Mandatory and Optional Properties," *Data and Knowledge Engineering* (55:3), pp 301-326.

Gregersen, H., and C. Jensen (1998) "Conceptual Modeling of Time-Varying Information," TR-35, TIMECENTER Technical Report, pp. 1-31.

Halpin, T.A. (2004) *Comparing Metamodels for ER, ORM and UML Data Models* Hershey, PA: Idea Group Publishing, pp. 23-44.

Hammer, M., and D. Mcleod (1981) "Database Description with SDM: A Semantic Database Model," *ACM Transactions on Database Systems* (6:3), pp 351-386.

Hardgrave, B.C., and N. Dalal (1995) "Comparing Object Oriented and Extended Entity Relationship Models," *Journal of Database Management* (6:3), pp 15-21.

Hermosilla, L.H. (1994) "A Unified Approach for Developing a Temporal GIS with Database and Reasoning Capabilities," 5th European Conference and Exhibition on Geographic Information Systems, EGIS '94, pp. 122-131.

Hull, R., and R. King (1987) "Semantic Database Modeling: Survey, Applications, and Research Issues," *ACM Computing Surveys* (19:3), pp 201-260.

Jensen, C.S., et al.(1998) "A Consensus Glossary of Temporal Database Concepts-February 1998 Version," in: O. Etzion, S. Jajodia and S. Sripada (eds.) *Temporal Databases: Research and Practice,* Springer-Verlag,

Juhn, S., and J.D. Naumann (1985) "The Effectiveness of Data Representation Characteristics on User Validation," International Conference on Information Systems, Indianapolis, IN, pp. 212-226.

Khatri, V., S. Ram and R.T. Snodgrass (2002) "Supporting User-defined Granularities and Indeterminacy in a Spatiotemporal Conceptual Model," *Annals of Mathematics and Artificial Intelligence* (36:1-2), pp 195-232.

Khatri, V., S. Ram and R.T. Snodgrass (2004) "Augmenting a Conceptual Model with Geospatiotemporal Annotations," *IEEE Transactions on Knowledge and Data Engineering* (16:11), pp 1324-1338.

Khatri, V., S. Ram, and R.T. Snodgrass (2006) "On Augmenting Database Design-Support Environments to Capture the Geo-Spatio-Temporal Data Semantics," *Information Systems* (31:2), pp 98-133

Kim, Y.-G., and S.E. March (1995) "Comparing Data Modeling Formalisms," *Communications of the ACM* (38:6), pp 103-113.

Korth, H., A. Silberschatz and S. Sudarshan (2005) *Database Systems Concepts*, (5th ed.) New York: McGraw Hill

Kramer, B., and Luqi (1991) "Towards Former Models of Software Engineering Processes," *Journal of Systems and Software* (15), pp 63-74.

Mantha, R.W. (1987) "Data Flow and Data Structure modeling for database requirements determination: a comparative study," *MIS Quarterly* (11:4), pp 531-545.

Mennecke, B.E., and M.D. Crossland (1996) "Geographic Information Systems: Applications and Research Opportunities for Information Systems Researchers," 29th Hawaii International Conference on System Sciences, pp. 537-546.

Moynihan, A. (1996) "An attempt to compare OO and functional decomposition in communicating information system functionality to users," Workshop on evaluation of modeling methods in systems analysis and design: CAiSE*96.

Mylopoulos, J., P. Bernstein and H. Wong (1980) "A Language Facilty for Designing Database Intensive Applications," *ACM Transactions on Database Systems* (5:2), pp 185-207.

Nijssen, G.M., and T.A. Halpin (1989) *Conceptual Schema and Relational Database Design: A fact-oriented approach*, Sydney, Australia: Prentice-Hall.

Palvia, P.C., C. Liao and P-L. To (1992) "The Impact of Conceptual Models on End-User Performance," *Journal of Database Management* (3:4), pp 4-15.

Parent, C., S. Spaccapietra and E. Zimányi (1999) "Spatio-temporal conceptual models: Data structures + space + time," 7th ACM Symposium on Advances in Geographic Information Systems, Kansas City, USA.

Peleg, M., and D. Dori (2000) "The Model Multiplicity Problem: Experimenting with Real Time Specification Methods," *IEEE Transactions on Software Engineering* (26:6), pp 1-18.

Ram, S. (1995) "Intelligent Database Design using the Unifying Semantic Model," *Information and Management* (29:4), pp 191-206.

Ram, S., and W. Wei (2003) "Semantic Modeling of Biological Sequences," 13th Annual Workshop on Information Technologies and Systems (WITS), Seattle, Washington.

Ram, S., and W. Wei (2004) "Modeling the Semantics of 3-D Protein Structures," Proceedings of the ER Conference.

Segev, A., and A. Shoshani (1987) "Logical Modeling of Temporal Data," Proceedings of the ACM SIGMOD, San Francisco, CA, pp. 454-466.

Seshadri, P., M. Livny and R. Ramakrishnan (1995) "SEQ: A Model for Sequence Databases," Proceedings of the International Conference on Data Engineering (ICDE),pp. 232-239

Shoval, P., and M. Even-Chaime (1987) "Database Schema Design: An Experimental Comparison Between Normalization and Information Analysis," *Database* (18:3), pp 30-39.

Shoval, P., and I. Frummerman (1994). "OO and EER Schemas: A Comparison of User Comprehension," *Journal of Database Management* (5:4), pp 28-38.

Siau, K. (1999) "Information Modeling and Method Engineering: A Psychological Perspective," *Journal of Database Management* (10:4), pp 28-38.

Siau, K. (2004) "Informational and Computational Equivalence in Comparing Information Modeling Methods," *Journal of Database Management* (15:1), pp 73-86.

Siau, K., J. Erickson and L. Lee (2005) "Theoretical versus Practical Complexity: The Case of UML," *Journal of Database Management* (16:3), pp 40-57.

Siau, K., Y. Wand, and I. Benbasat (1997) "The Relative Importance of Structural Constraints and Surface Semantics in Information Modeling," *Information Systems* (22:2 and 3), pp 155-170.

Smith, J.M., and D.C.P. Smith (1977) "Database Abstractions: Aggregation and Generalization," *ACM Transactions on Database Systems* (2:2), pp 105-133.

Snodgrass, R.T. (1987) "The Temporal Query Language TQuel," *ACM Transactions of Database Systems* (12:2), pp 247-298.

Snodgrass, R.T. (1999) *Developing Time-Oriented Database Applications in SQL* Morgan Kaufmann Series in Data Management Systems, San Francisco, p. 504.

Snodgrass, R.T., and I. Ahn (1986) "Temporal Databases," *IEEE Computer* (19:9), pp 35-42.

Snodgrass, R.T., et. al (1997) "Transitioning Temporal Support in TSQL2 to SQL3," in: O. Etzion, S. Jajodia and S.M. Sripada (eds.),*Temporal Databases: Research and Practice,* Springer Verlag, New York, pp. 150-194.

Topi, H., and V. Ramesh (2002) "Human Factors on Research on Data Modeling: A Review of Prior Research, An Extended Framework and Future Research Directions," *Journal of Database Management* (13:2), pp 3-19.

Tryfona, N. and C.S. Jensen (1999) "Conceptual Data Modeling for Spatiotemporal Applications," *Geoinformatica* (3:3), pp 245-268.

Tsichritzis, D.C. and F.H. Lochovsky (1976) "Hierarchical Database Management: A Survey," *Computing Surveys*:March), pp 105-124.

Verhijen, G.M.A., and J.V. Bekkum (1982) "NIAM: An Information Analysis Method," IFIP WG 8.2 Conference on Comparative Review of Information System Design Methodologies, New York, NY: North-Holland.

Wand, Y., and R. Weber (1993) "On the Ontological Expressiveness of Information Systems Analysis and Design Grammars," *Journal of Information Systems* (3), pp 217-237.

Wand, Y., and R. Weber (1995) "On the Deep Structure of Information Systems," *Information Systems Journal* (5), pp 203-223.

Wand, Y., and R. Weber (2002) "Information Systems and Conceptual Modeling: A Research Agenda," *Information Systems Research* (13:4), pp 363-376.

Woodsmall, R.M., and D.A. Benson (1993) "Information Resources at the National Center for Biotechnology Information," *Bulletin of the Medical Library Association* (81:3), pp 282-284.

Worboys, M.F., H.M. Hearshaw, and D.J. Maguire (1990) "Object-oriented data modeling for spatial databases," *International Journal of Geographical Information Systems* (4:4), pp 369-383.

ABOUT THE AUTHORS

Gove N. Allen holds bachelors and masters degrees in Accountancy from Brigham Young University. Dr. Allen has consulted for many major corporations on the implementation of database technology including Sony, AT&T, Sprint, Hewlett Packard, Micron, Intel, 3M, American Express, and the Kennedy Space Center. More recently, he developed WebSQL.org, a site for teaching database management that allows dynamic execution of Structured Query Language against Oracle databases through a simple web interface. Dr. Allen received his Ph.D. from the University of Minnesota in 2001 and currently serves as an assistant professor of e-business and information systems at Tulane University's A. B. Freeman School of Business.

Akhilesh Bajaj is Chapman Associate Professor of MIS, at the University of Tulsa. He received a B. Tech. in Chemical Engineering from the Indian Institute of Technology, Bombay in 1989, an MBA from Cornell University in 1991, and a Ph.D. in MIS (minor in Computer Science) from the University of Arizona in 1997. Dr. Bajaj's research deals with the construction and testing of tools and methodologies that facilitate the construction of large organizational systems, as well as studying the decision models of the actual consumers of these information systems. He has published articles in several academic journals such as *Management Science, IEEE Transactions on Knowledge and Data Engineering, Information Systems* and the *Journal of the Association of*

Information Systems. He is on the editorial board of several journals in the MIS area. His research has been funded by the department of defense (DOD). He teaches graduate courses on basic and advanced database systems, management of information systems, and enterprise wide systems.

Vijay Khatri is an Assistant Professor of Information Systems in the Kelley School of Business at Indiana University. He received his Ph.D. in Management Information Systems from the University of Arizona. His research interests center on developing techniques to facilitate elicitation of spatial/temporal database requirements and understanding the cognitive aspects that influence analysis and design of information systems. His work has been published or is forthcoming in Information Systems, IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Professional Communication, IEEE Transactions on Education, and Annals of Mathematics and Artificial Intelligence, among others. He is a member of IEEE, the Association of Information Systems (AIS), ACM and Informs.

Sudha Ram is Eller Professor, Management Information Systems in the Eller School of Management at the University of Arizona. She received a B.S. degree in mathematics, physics and chemistry from the University of Madras in 1979, PGDM from the Indian Institute of Management, Calcutta in 1981 and a Ph.D. from the University of Illinois at Urbana-Champaign, in 1985. Dr. Ram has published articles in such journals as Communications of the ACM, IEEE Expert, IEEE Transactions on Knowledge and Data Engineering, Information Systems, Information Systems Research, Management Science, and MIS Quarterly. Dr. Ram's research deals with issues related to Enterprise Data Management. Her research has been funded by organizations such as, IBM, Intel Corporation, Raytheon, US ARMY, NIST, NSF, NASA, and Office of Research and Development of the CIA. Specifically, her research deals with Interoperability among Heterogeneous Database Systems, Semantic Modeling, BioInformatics and Spatio-Temporal Semantics, Business Rules Modeling, Web services Discovery and Selection, and Automated software tools for database design She is also the director of the Advanced Database Research Group based at the University of Arizona.

Keng Siau is a Professor of Management Information Systems (MIS) at the University of Nebraska-Lincoln (UNL). He is currently serving as the Editor-in-Chief of the *Journal of Database Management* and as the Book Series Editor for *Advances in Database Research*. Dr. Siau has published more than 80 refereed journal articles, and these articles have appeared in journals such as MISQ, CACM, IEEE Computer, Information Systems, IEEE Transactions on Systems, Man, and Cybernetics, IEEE Transactions on Professional Communication, IEEE Transactions on Information Technology in Biomedicine, IEEE Transactions on Education, and others. He served as the Organizing and Program Co-chairs of the *International Workshop on Evaluation of Modeling Methods in Systems Analysis and Design (EMMSAD)* (1996–2005). For more information, please visit his website at URL: http://www.ait.unl.edu/siau/

Acknowledgment: The authors thank the Editor-in-Chief of *CAIS* and the attendees of the panel at AMCIS 2005. Their comments and feedback helped improve the quality of this paper.

.

Communications of the Association for Information Systems

ISSN: 1529-3181