A Taxonomy of Data Modeling Techniques

Vladan Jovanovic
vladan@georgiasouthern.edu

Adrian Gardiner

Ljiljana Cupic

Follow this and additional works at: http://aisel.aisnet.org/sais2007

Recommended Citation
http://aisel.aisnet.org/sais2007/34

This material is brought to you by the Southern (SAIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in SAIS 2007 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
A TAXONOMY OF DATA MODELING TECHNIQUES

Vladan Jovanovic  
Georgia Southern University  
vladan@georgiasouthern.edu

Adrian Gardiner  
Georgia Southern University  
tagardine@georgiasouthern.edu

Ljiljana Cupic  
Georgia Southern University  
lilycj48@yahoo.com

Abstract

This paper presents a proposed taxonomy of data modeling techniques. Each technique was classified with regards to its primary source of domain knowledge, recommended/intended use with regards to system size, and whether the technique was analytical or more synthesis oriented. We predict that our taxonomy will prove valuable to both academics and practitioners, and form the basis of future endeavors aimed at developing more robust methodologies for developing standardized data models (see Appendix B), large data models, and applications of ontologies to real-world database design.

Keywords: Data modeling, data model patterns, data base design, ontology

Introduction

This paper presents a preliminary report from a larger study (ongoing) that is investigating issues in the development of domain-dependent standardized data models and related opportunities for ontology driven data modeling. The larger study is motivated by a desire to speed up development and integration processes through reuse of knowledge and expertise captured by such models, as well as to develop a viable methodology for creating large models within new domains. Thus, our primary interest is to improve modeling practice for large systems, as few initiatives, as reported by (Moody, 2005), have dealt with properties pertinent to large scale models. We should, however, acknowledge that using ontologies to lead data modeling efforts on very large systems has not been mentioned in recent work on applying ontologies for data modeling (Krogstie et al., 2005; Sugumaran & Storey, 2005). Specifically, the current paper presents a taxonomy (classification) of data modeling techniques.

Large systems and data models

There is no well-recognized definition of what constitutes a large system. Such a discussion raises numerous ontological issues, such as the mereological problem of what constitutes the boundary of a system. One choice is to define a large system by focusing upon correlates of system size. For example, one would expect a non-linear monotonic relationship between system size and both system and structural complexity. On this last point, Milling (2002) suggests there are three broad dimensions of complexity: the number of relevant elements (quantity), the number of connections between elements (connectivity), and functional intricacy of the connections between the elements (functionality). In addition, a number of metrics of system complexity have been proffered, such as the depth of inheritance trees (Chidamber & Kemerer, 1994). However, as Mayer & Hall (1999, p.103) warns in terms of OO systems: “The term complexity is commonly used without any clear definition being given of its meaning...” Mayer & Hall (1999) also criticize existing metrics of system complexity as being unclear.
As an alternate approach, one can apply indirect measures of system size, such as the amount of work required to construct it (i.e., rather than use a structural approach to defining system size, use an input approach). In this respect, we posit a preliminary definition of a large system as one that has required an investment in excess of 100,000 labor hours. This definition is not without a problem, as frequently evidence of effort is hard to verify and compare from system to system. Nevertheless on suspecting/determining the existence of large system size, one can focus on measuring associated model size. For this purpose, one acceptable metric is the number of featured terms (as number of distinct entity types, relationships, and attributes, or their equivalents) within a model representation. The reason selection of a model size measure is that the number of terms represents, in fact, the minimal number of definitions needed to be covered within a model. Moreover, we will adopt an ordinal scale of model size, and will define medium size models to refer to models that have between 100 and 1,000 terms, for large model size interval between 1,000 and 10,000 defining terms was chosen, while small and very large are referring to models with less then 100 or over 10,000 terms, respectively. Although one can claim these numbers are arbitrary, they conveniently support our intuition of a very steep increase of efforts with size and permit use of logarithmic scale in comparisons. Given this classification of sizes most of the models found in textbooks and research papers address the least interesting and, from the standpoint of techniques used, the least demanding group of small systems.

Data modeling standards

The subject of data modeling is treated here very broadly and covers conceptual modeling and, in particular, standardized enterprise/domain modeling using various notations. Among the most prevalent notations in literature (Appendix A) are Entity Relationship (ER) based ones, started with the original ER (Chen, 1976), like IDEF1X (Bruce, 1992), Oracle CASE notation (Barker, 1989; Hay, 1996), IE (Simsion, 2005), and other variations (too numerous to quote). Also in extensive use worldwide are approaches and specialized notations competing with ER, such as EXPRESS (Schenk & Wilson, 1994), RM-ODP (Kilov, 1999), and ORM (Halpin, 2001). Within the scope of our research are also a number of alternative object modeling approaches and notations (e.g., UML, OMT, Coad-Yourdon, Shlaer-Mellor object models, etc.). See Appendix A); and other data structuring notations, including XML Schema Design Diagrams (Kim, 2003), dimensional modeling diagrams (Kimbal, 2002) and experimental approaches in contemporary service-based business and system specifications, like Coordination-Contracts (Andrade & Fiadero, 2001). Of the modeling conventions mentioned, only four: EXPRESS (ISO, 1994), IDEF1X (IEEE, 1998), RM-OMP (ISO/IEC, 2000), and UML (ISO/IEC, 2004) have been formally balloted in an open process and designated as standards.

Practitioners and academicians on data modeling

Part of our motivation for writing this paper is to address a perceived gap between the viewpoints of practitioners and academics in the area of data modeling (Batra & Marakas, 1995), as well as confusion over an appropriate research agenda for conceptual modeling (Wand & Weber, 2002). For example, Morein (2006) recently concluded that Data Modeling textbooks are in a confused, ambiguous and contradictory state. A specific criticism from Morein (2006) is that he feels many academics who teach data modeling may not adequately emphasize or appreciate many real-world issues and problems faced by practitioners. Further comment on this alienation comes from Simson (2005), who positions practitioners in the design camp, and academics in the analysis camp. It is therefore our intent that the taxonomy of data modeling techniques presented in this paper will help bridge some of this divide by presenting a more balanced viewpoint – through situating analytical techniques, used mainly for small systems, within the use of more direct design techniques, which we will assert are more appropriate for larger systems. However, before going any further, we will attempt to clarify our use of the terms: ‘data modeling’, and ‘technique.’

Data modeling is the activity of creating a data model, in the sense of a schema as per (Hirshheim et al., 1995). Hirshheim et al.’s definition is used to cover activities of data semantics discovery and/or definition leading to creating or evolving data schemas including the XML schemas. In contrast, the term ‘technique’ applies freely to any practical approach that is recognized or can be explained to practitioners without much ado. In this regard, the study of practice is prima facie the researching of consequence, and we felt more comfortable, in the preliminary stage of our research, to adopt a broader and the less ambitious term ‘technique,’ thus allowing us to discriminate effectively without forcing definitional burden other than clear delineation of one technique from the other. The reason why we use the term ‘technique,’ rather than the more scientific ‘method,’ lies in the fact that data modeling is not a science, but a practice, in the same sense that scientific research is not itself a science but is still an art or...
craft similarly practiced by educated, knowledgeable, experienced and dedicated humans, that is practitioners as professionals.

Categorizing data modeling techniques

Our taxonomy emerged as a two level categorical classification, using strategic choice of analytic vs. synthetic approach as the first level criterion, and a tactical choice of information sources as the second level criterion. From the standpoint of which primary approach to knowledge discovery is used within the data modeling process, we have identified and named the two distinct categories of techniques: a) identification techniques, and b) direct modeling techniques. This first level of classification discriminates along the analytic vs. synthetic (design) dichotomy of knowledge use. The two general approaches are also referred to as bottom-up and top-down, the terms in common use within academic community.

Identification techniques

Techniques falling into this category are more strategic. We chose the term identification as it clearly indicates that externally visible evidence is used as basis for analysis. All identification-based techniques listed in the Table 1 focus on parsing and restructuring given samples and or definitions of data sets, including internal schemas, in the terminology of three-level ANSI/SPARC architecture. The outcomes of identification techniques are data models, or more likely view-based partial data models.

Frequently, work on producing data models using identification techniques can be accomplished with tools and/or by inexperienced designers and hence the attraction. The last point may be the root cause of the broad adoption of identification (bottom-up) techniques in teaching as well as of a sustained research interest in formalizing, automating and promoting such techniques. Given the aura of exactness and rigor, an obvious and well known example here is normalization- featured in practically every database textbook. One expectation was that automation will scale the techniques to sizable systems, but after 20+ years of research and development we do not have evidence of success in that. The practice on the other hand, favor experience as witnessed in emergence of patterns, use of standard models, and standardized applications (embodying complex data models as SAP does) as well as in a sustained development of tools for computer aided design. Professional association provided a bibliography (DAMA, 2001) that agrees (over 90% overlap) with a much larger set of 100 professional sources used in our study (Appendix A), indicating no interest in identification based automated methods except reverse engineering. A prerequisite for reverse engineering is an existing schema, and the problem of how to come up with the schema (model) in the first place remains unsolved.

Direct modeling techniques

The techniques within the direct modeling group are listed in Table 2. The primary differentiator among these techniques is, again, the form of knowledge representation to be used - but this time, clearly it is the knowledge of the overall system itself and not the data content of its external manifestations that is the driver. The case of data mining as a basis for data modeling is clearly peculiar. The critical role of data models in preparing for data mining is well known: a Google search on ‘data modeling for data mining’ returned numerous references to mature industrial use. But that is not the subject of our speculation here. We are experimenting with data mining as a tool to hypothesize potentially relevant relationships and, as such opening interesting insights into complex systems. So, it may seem that this category could have been listed among identification techniques; we subscribe to the point of view that human insight will be the driver, not the rote computation in eventual uses of Data Mining for Data Modeling.

The last two techniques in Table 2 are based on structuring and evolution of data models aimed at up-front synthesis in order to assure shared core model and avoid excessive view integration problems common in models developed using identification techniques. The last category is of more recent origin, for an overview relating data modeling and ontology development and a related ontology focused approach for databases see respectively (Gasevic et al., 2006) and (Sugumaran & Storey, 2005). With a recent surge of research re semantic web systems and models, the associated techniques considering ontology are under intense development (Sugumaran & Storey, 2005; Wand & Weber (2002); Krogstie et al., 2005) and we expect more techniques, besides ours, in this category to emerge soon. A caveat is in order, while the number of direct modeling techniques was developed and supported with tools under

Proceedings of the 2007 Southern Association for Information Systems Conference 191
clear understanding of limits of such automation, we are facing somewhat uncritical embrace of semantic web and XML ‘technologies’ as another ‘silver bullet’ (Fensel, 2005).

Table 1. Taxonomy of Data Modeling Techniques: Identification Category

<table>
<thead>
<tr>
<th>Technique</th>
<th>Definition</th>
<th>Recommended usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity expansion</td>
<td>Identifying subjects in existing data stores, file descriptions etc. and reassigning attributes to new entities</td>
<td>Small scale models or in support of medium sized ones. See for example (Model, 1992; Bruce 1992).</td>
</tr>
<tr>
<td>Input Data Analysis-structuring</td>
<td>Transforming data entry forms, screens, and questionnaires into views</td>
<td>Isolated small applications or in support of medium and large models. Featured in textbooks, see also (Jovanovic &amp; Mrdalj, 1990).</td>
</tr>
<tr>
<td>Output Data Analysis - structuring</td>
<td>Transforming representative documents, reports, query results, XML documents/schemas etc. into views</td>
<td>Typically multiple outputs are analyzed and integrated for small to medium models; also recommended to be used for XML schema design. Frequently used and well represented in textbooks, also (Barker, 1989).</td>
</tr>
<tr>
<td>Normalization</td>
<td>Partitioning given relations based on known (semantics) dependencies</td>
<td>Generally not used as a design method but rather a quality control one, (Harrington, 2002; Hernandez, 2003; Hobermans, 2005).</td>
</tr>
<tr>
<td>Schema Reverse Engineering</td>
<td>Mapping of relations, primary and foreign keys and other attributes to data models entities, relationships and attributes.</td>
<td>Used with aid of software tools in cases of porting and integrating medium sized systems. For discussion see for example (Allen, 2002; Ambler &amp; Sadalage, 2006).</td>
</tr>
</tbody>
</table>

Table 2. Taxonomy of Data Modeling Techniques: Direct Modeling Category

<table>
<thead>
<tr>
<th>Technique</th>
<th>Definition</th>
<th>Recommended usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert knowledge</td>
<td>Expert knowledge representation from scratch or elaboration from existing data models</td>
<td>Useful for specialized or individualized small to medium-sized systems. For a discussion see (Simsion &amp; Witt, 2005).</td>
</tr>
<tr>
<td>Text analysis</td>
<td>Analysis of system descriptions</td>
<td>Small systems (Appendix A).</td>
</tr>
<tr>
<td>Data mining</td>
<td>Discerning unsuspected relationships for analysis</td>
<td>At this point is only speculative, possibly beneficial as supporting technique where large data sets are available – no definitive sources in public domain are readily available.</td>
</tr>
<tr>
<td>Patterns or standard models</td>
<td>Adaptation of patterns or standardized models relevant for a domain of inquiry</td>
<td>Medium to large sized domains. For examples see (Gessford 1991; Hay, 1996; Fowler, 1997; Silverston, 2001; Jones &amp; Song, 2005).</td>
</tr>
<tr>
<td>Ontology-based modeling</td>
<td>Data Modeling starts with ontology and evolves sub-models in pre integrated fashion.</td>
<td>Intended for large systems; used by authors since 90’s. Relevant source addressing large scale models is (Scheer, 1989).</td>
</tr>
</tbody>
</table>

Conclusion

This paper presents a proposed taxonomy of data modeling techniques. In addition, we made explicit for each technique within the taxonomy its primary source of domain knowledge and recommended/intended use with regards to system size. Although we initially limited our taxonomy to ten distinct techniques, we acknowledge that these techniques may be applied with other ad hoc and/or proprietary approaches that are at the modeler’s disposal. In terms of intended use, we claim that only the last two techniques included in Table 2 have the potential to scale to very large systems. However, to support this last assertion and eventually formulate guidance for practitioners, there is a need for the development of a formal model for analysis, which we intend to address in future research. Finally, directions for possible future study can be summarized as follows: a) perform detailed semantic analysis of large models (Appendix B) looking for underlining ontology patterns and application of concepts, b) elaboration of ontology-based modeling technique, and c) the eventual formulation of a viable unifying method that can help model databases, data warehouses and also domain ontologies.
References


Proceedings of the 2007 Southern Association for Information Systems Conference 193


**Appendix A. Annotated bibliography: data modeling**

Annotated bibliography of over 100 professional books and over 30 relevant textbooks: e-mail to obtain.

**Appendix B. Standardized data models** (sample only)

1. ACCORD Data Model for Insurance (for members only).
7. Public Petroleum Data Model (PPDM), http://www.ppdm.org
10. Unified POS, Retail Standard Data Model, 2005 (note: for members only; XML Schema is free).