Information & Management 49 (2012) 151–163

Contents lists available at SciVerse ScienceDirect

Information & Management

journal homepage: www.elsevier.com/locate/im

Data modeling: Description or design?

Graeme Simsion, Simon K. Milton *, Graeme Shanks

Department of Computing and Information Systems, The University of Melbourne, Victoria 3010, Australia

ARTICLE INFO

Article history:
Received 13 November 2010
Received in revised form 15 November 2011
Accepted 25 January 2012
Available online 14 February 2012

Keywords:
Conceptual data modeling
Practitioner study
Design
Analysis

ABSTRACT

Data modeling for database creation has generally been considered to be a *descriptive* process: the real-world is observed and represented in a conceptual model that is then transformed into a logical structure for a database. This is reflected in prescriptive methods and is the dominant assumption in most studies. However, data modeling can also be considered a type of *design* with negotiable requirements, a creative process, and many workable solutions. Our paper discusses empirical results from almost 500 practitioners on three continents comparing data modeling to *design*. We found that data modeling, as practiced, was better characterized as design.

1. Introduction

1.1. Alternative views of data modeling

Data modeling is one of the most critical activities in the implementation of an IS: it has been characterized as a process of *reality mapping*. This characterization has been occasionally challenged from a philosophical perspective, from observations of practice, and from empirical evidence.

This descriptive characterization also dominates the practitioner literature. In a descriptive activity, a set of artifacts may be created, and this might well be called design, but not be of sufficient importance to the overall result as to characterize the entire activity as design. In data modeling, there is choice in the selection of components (typically entities, relationships and attributes) used to represent some part of reality. The difference between description and design is in whether this selection is a *trivial part* of the process compared to understanding the Universe of Discourse (UoD) {descriptive type}, or whether it is the essence of the process {design type}.

1.2. Previous empirical research

We know little about how experienced data modelers approach their work or about the models that are produced for real business applications. This is because most studies have assumed the process to be descriptive and have not involved practitioners.

The descriptive characterization is embodied in most empirical studies through the use of a *gold standard* – a single correct solution devised by the researcher, who often embedded entity and relationship names in descriptions, thus constraining the modeling abstractions. The "business requirements" amounted to a plain language description and the participant's task was to translate the description to the original diagram. For example, "An employee can report to only one department. Each department has a phone number." Tasks showing these two traits mainly tested facility with modeling formalisms. Yet it is common to see conclusions that indicated that novice designers did not run into much trouble in modeling entities and attributes. In the context of the research task, modeling entities may have meant little more than identifying nouns in the description. The use of simple models and prescriptive instructions limited the scope of the design.

Most empirical studies have used students as participants; of course, this limited the difficulty of the problems posed. Of the total of 3210 participants across 59 studies that we surveyed, only 147 in nine studies had more than one year's industry experience of data modeling. Thus most studies used unrealistically simple data models. Nevertheless, some studies have used experienced data modelers and have uncovered design behavior. Comparisons of novice and expert data modelers have revealed behaviors characteristic of designers (attempting to gain a holistic understanding, categorization of problems, pattern re-use) in the experts but not in the novices.

1.3. The research question

Our research question was: *Is data modeling better characterized as description or design?* Here, *data model* refers to a model of a specific UoD (e.g. the data model of ABC corporation's human

* Corresponding author. Fax: +61 3 9349 4596.
E-mail address: simon.milton@unimelb.edu.au (S.K. Milton).

0378-7206/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.im.2012.01.003

resources operations), while *data modeling* refers to the set of activities required to specify a *conceptual schema* that will transform into a database schema but *prior* to its transformation into a specific DBMS data definition.

Specifically, we examined the process of data modeling that resulted in a database design that could be implemented in a relational DBMS. We did not include other purposes of conceptual data modeling (e.g., its use in IS planning).

This is an important research question for at least three reasons:

- Most data modeling research assumes the descriptive characterization, notably in the design of experiments and in the application of ontology [2,3]. If data modeling is, in fact, design, research results need to be reinterpreted in that light.
- Data modeling education should include expert level practice.
- Creative thinking and evaluation of alternative designs are intrinsic to design processes. If data modeling is seen as a design activity rather than description, then data modeling methods should be updated to reflect a design process and explicitly include creative thinking and comparative evaluation.

2. The ideal type of 'design'

Design is the *ideal type* against which we measure data modeling. Its essence has been synthesized in the form of a list of some of the important characteristics of design problems and solutions, and the design process itself. These characteristics are intended to *typify* design and thus to *differentiate* it from description.

The list is not exhaustive, and the characteristics are interrelated. Collectively, they provide an overall picture of design. The characteristics (shown in Table 1) were grouped into fourteen properties within three dimensions – *Problems*, *Solutions* (i.e., *Products*), and *Process*, following Lawson's [1] properties of design

Table 1
Lawson's properties of design.

Design problems
1. Design problems cannot be comprehensively stated
2. Design problems require subjective interpretation
3. Design problems tend to be organized hierarchically
The design process
1. The process is endless
2. There is no infallibly correct process
3. The process involves finding as well as solving problems (including creativity)
4. Design inevitably involves subjective value judgments
5. Design is a prescriptive activity
6. Designers work in the context of a need for action
Design solutions
1. There are [sic] an inexhaustible number of different solutions
2. There are no optimal solutions to design problems
3. Design solutions are often holistic responses
4. Design solutions are a contribution to knowledge
5. Design solutions are parts of other design problems

3. Research design

The framework provides a basis for expanding the research question to 11 research sub-questions (RSQs) (see Table 2). *Scope* seeks to clarify what practitioners mean by data modeling. The remaining three dimensions determine whether data modeling practice has the properties of the design type: *Problem* deals with the negotiability of data modeling requirements, *Process* examines whether data modeling is creative, and *Product* deals with the diversity in data models produced by experienced practitioners in response to a task; this suggests that data modeling is a design process.

The RSQs were addressed with a combination of surveys, laboratory studies, and interviews. The selection of the mode for each is shown in Table 3. Semi-structured interviews (with influencers of practitioners or *thought leaders*), surveys (to collect the perceptions of experienced data modelers about data modeling products, processes, and problems), and laboratory studies (designed to explore diversity and style in data models by asking participants to complete modeling tasks which were examined for evidence of diversity, style, and patterns in data modeling) were chosen to provide multiple sources of data to assess the practice of data model against the ideal type of design.

The surveys and laboratory studies were incorporated into 12 data modeling seminars and workshops for experienced practitioners delivered by the first author in the US, UK, Scandinavia and Australia between May 2002 and November 2004 (see Appendix A for summary of participants). Figs. 1–4 summarize the responses to demographic questions.

There was a strong correlation between the two experience measures ($\gamma = 0.65$, $p < 0.0005$). Our study was the largest currently published; it included 381 participants with at least one year of data modeling experience. The minimum number to complete any task was 55. Three other groups participated in the research. The *Practitioner thought-leaders* and *expert model evaluators* were purposive samples. *Architects and accountants* (who provided a benchmark for the *Characteristics of Data Modeling* component) were recruited from personal and professional contact lists.

Table 2
Research sub-questions (RSQs).

Type dimension	Name	Research sub-question
General – applying to all three dimensions	Scope	(RSQ1) What do data modeling practitioners believe is the scope and role of data modeling within the database design process?
	Importance	(RSQ2) Is the description/design question considered important by data modeling practitioners?
	Espoused Beliefs	(RSQ3) What are the (espoused) beliefs of data modeling practitioners on the description/design question?
Problem	Perception of Problems	(RSQ4) Are data modeling problems perceived as design problems by data modeling practitioners?
Process	Methods	(RSQ5) Do database design methods used in practice support a descriptive or design characterization of data modeling?
	Perception of Processes	(RSQ6) Are data modeling processes perceived as design processes by data modeling practitioners?
Product	Perception of Products	(RSQ7) Are data modeling products perceived as design products by data modeling practitioners?
	Diversity in Conceptual Modeling	(RSQ8) Will different data modeling practitioners produce different conceptual data models for the same scenario?
	Diversity in Logical Modeling	(RSQ9) Will different data modeling practitioners produce different logical data models from the same conceptual model?
	Patterns	(RSQ10) Do data modeling practitioners use patterns when developing models?
	Style	(RSQ11) Do data modeling practitioners exhibit personal styles that can be identified in the data models that they create?