Available online at www.sciencedirect.com
ScienceDirect
Decision Support Systems 42 (2006) 1170–1186
Decision Support Systems
www.elsevier.com/locate/dsw

# Selecting optimal instantiations of data models—Theory and validation of an ex ante approach

P.L. Bowen [a], R. Debreceny [b], F.H. Rohde [a,*], J. Basford [a]

[a] University of Queensland, UQ Business School, 4072, Australia
[b] University of Hawaii, College of Business Administration, Honolulu, 96822, USA

Received 20 July 2004; received in revised form 7 October 2005; accepted 13 October 2005
Available online 4 January 2006

**Abstract**

The schema of an information system can significantly impact the ability of end users to efficiently and effectively retrieve the information they need. Obtaining quickly the appropriate data increases the likelihood that an organization will make good decisions and respond adeptly to challenges. This research presents and validates a methodology for evaluating, ex ante, the relative desirability of alternative instantiations of a model of data. In contrast to prior research, each instantiation is based on a different formal theory. This research theorizes that the instantiation that yields the lowest weighted average query complexity for a representative sample of information requests is the most desirable instantiation for end-user queries. The theory was validated by an experiment that compared end-user performance using an instantiation of a data structure based on the relational model of data with performance using the corresponding instantiation of the data structure based on the object-relational model of data. Complexity was measured using three different Halstead metrics: program length, difficulty, and effort. For a representative sample of queries, the average complexity using each instantiation was calculated. As theorized, end users querying the instantiation with the lower average complexity made fewer semantic errors, i.e., were more effective at composing queries.


*Keywords:* Models of data; Data representations; Object-relational databases; Relational databases; Query languages; Query complexity

## 1. Introduction

Over the last decade many organizations have expanded their transaction processing databases, implemented enterprise resource planning (ERP) systems, and built enterprise-wide data warehouses. Maximizing their returns from these investments requires organizations to make these data repositories available directly to knowledge workers for operational, tactical, and strategic decision making [48]. Knowledge workers can access these data repositories via a wide variety of end-user analytical tools including graphical query interfaces, report writers, OLAP cube builders, and data mining tools as well as the more traditional database query languages [44,46].

The value of the analyses made by knowledge workers depends on the quality of the information captured by *and* retrieved from the enterprise's data repositories. Obviously, the accuracy of the data retrieved is affected by the accuracy of the data stored. One hundred percent accurate stored data does not, however, guarantee that

* Corresponding author. Tel.: +61 7 3365 6530; fax: +61 7 3365 6788.
*E-mail address:* f.rohde@business.uq.edu.au (F.H. Rohde).

0167-9236/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2005.10.002

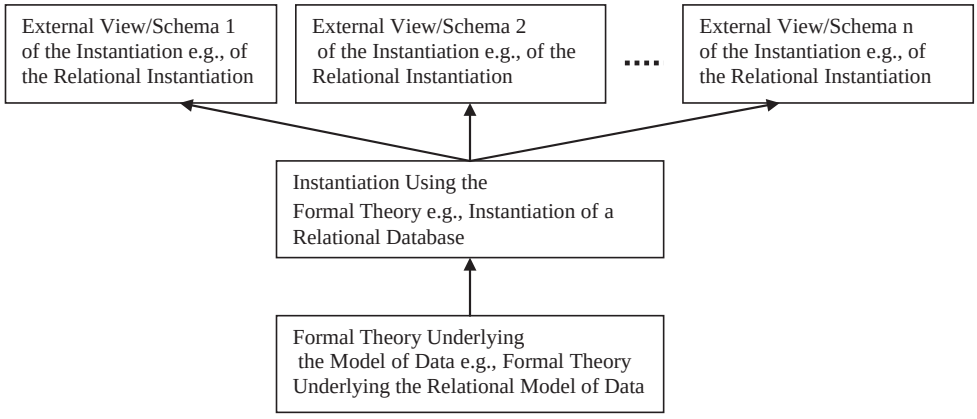


Fig. 1. General hierarchy of theoretical models of data, instantiations, and external views.

the data retrieved will be accurate. The accuracy of the data retrieved also depends upon the semantic[1] accuracy of the end-user composed queries.

This research presents and validates a methodology for evaluating, ex ante, the relative desirability of alternative instantiations of a model of data. In contrast to prior research, each instantiation is based on a different formal theory. A database is an instantiation of a formal theory of a model of data. Different sets of end users may be presented with different representations (external views/schemas) of the same instantiation. End users update data and retrieve information from the instantiation of one or more of the external schemas. These relationships are depicted in Fig. 1. An application of this general notion to a specific formal theory, i.e., relational database theory, is also depicted in Fig. 1. A relational database system is "based on a formal theory called the relational model of data" [15, pg. 15]. Thus, when the organization develops a database using a relational database management system, they are developing an instantiation based on the formal theory of "the relational model of data." Once the organization's instantiation of the relational database has been created, the organization must then decide which different representations (external views/schemas) of the relational instantiation to provide to the end users. For example, the organization may provide a set of third normal form (3NF) schemas for data capture and a set of first normal form (1NF) schemas for data retrieval.

Basford et al. [3] investigated the effect of providing end users with different external schemas based on the same instantiation of a database. In particular, they empirically tested the theory that, when faced with multiple possible data representations, end users perform information retrieval tasks better using the data representation that yields queries with lower average complexity. Basford et al. [3] used two external schemas: one in 1NF and another one in 3NF. Prior to conducting their experiment, they constructed a representative set of queries and, ex ante, evaluated the average complexity for the query formulations in each external schema. The 1NF schema produced, on average, less complex queries. As predicted by the theory, Ref. [3] found that end users who were given an external schema that, ex ante, led to less complex queries performed better.

Organizations are constantly faced with deciding which external schemas to offer to their end users. Less frequently, organizations are faced with the more difficult and more important task of evaluating different instantiations of databases based upon different formal theories of models of data. This paper examines whether the algorithm tested by Ref. [3] to optimize external schemas based on the same underlying model of data holds when comparing instantiations created via different formal theories of models of data.

This paper extends prior research in the following ways. First, this paper increases the precision and formalism of the methodology advocated to select from competing schemas. Second, Ref. [3] provided an initial test of the theory when optimizing two external schemas/views, a 1NF schema versus a 3NF schema, derived from the same instantiation of a relational

[1] Syntactic errors are errors in a programming statement that are due to incorrect use of the grammar of the language. Here, the reserved words and symbols are the SQL (Structured Query Language) operators and operands. Semantic errors are errors in a programming language statement such that the statement performs a function that differs from the one the user intended. In SQL, a semantic error occurs when the query does not correspond to the meaning of the user's request.