Informatics in Education, 2003, Vol. 2, No. 2, 191–200

An Empirical Perspective of Using Ternary Relationships in Database Conceptual Modelling

Elena CASTRO, Dolores CUADRA, Paloma MARTÍNEZ
Computer Science Department, Escuela Politécnica Superior
Universidad Carlos III de Madrid
Avda. de la Universidad 30, 28911 Leganés Madrid, Spain
e-mail: {ecastro, dcuadra, pmf}@inf.uc3m.es

Received: March 2003

Abstract. To apply Extended Entity Relationship Model (EER) is a good method for representing requirements on information systems, because of its high level of abstraction. Although it is very close to the user, it is not so trivial when some constructs, such as higher order relationships, are used. This paper describes the characterisation and several important results of an experiment performed at our university in order to show some of the difficulties found when novice students and practitioners use ternary relationships. Some special topics in identifying ternary relationships such as the importance of the domain of text and the intersection data are also investigated. In order to guide and help users in the design task, these results are introduced in PANDORA Case Tool, a research project which tries to serve as a methodological assistance tool.

Key words: CASE tools, database design methodologies, intelligent tutoring systems.

1. Introduction

EER model is well known as one of the best conceptual models in representing data requirements within database design, because of simplicity and easy conceptualisation (Hitchman, 1995). Nevertheless, some semantic constructs are not well understood for novice and practitioners and are also difficult to be detected by expert designers. In that sense, several studies have been developed for providing heuristics to make easier the modelling work (Batra and Antony, 1994). Chen's first research (Chen, 1976) proposed certain rules in order to detect constructs; for instance, an entity can be detected by a substantive, an attribute by an adjective and a relationship by a verb. Batra (Batra and Zanakis, 1994) states that "Past research has shown that non-experts have little or no difficulty in modelling entities and attributes, but considerable problems in modelling relationships".

Two different matters can be arisen in detecting relationships:

1. A relationship is a more complicated semantic concept than others due to its characteristics: degree (number of entities that participate in a relationship), connectivity (mapping among the instances of each entity) and cardinality constraints (maximum and minimum number of instances in one entity that can be associated in a relationship to a single instance in other entity) (Martinez *et al.*, 2000).

2. There is a lack of clarity in describing this construct in some textual specifications together with the ignorance about certain domains.

In general, case studies developed about difficulties in modelling relationships have been focused on the scope of binary relationships, particularly, on the degree and connectivity properties (Hitchman, 1995; Goldstein and Storey, 1990; Siau *et al.*, 1995; Batra and Davis, 1992). Only certain works in ternary relationship modelling have been developed (Batra and Antony, 1994; Batra and Zanakis, 1994; Bock and Ryan, 1996) and even some authors do not take into account higher order relationships[1] because of their low application frequency (Martinez *et al.*, 1999).

In this research we study the behaviour of novice users and practitioners in modelling ternary relationships. For instance, we are interested in the clues existing in textual descriptions that could help to discover such constructs and the difficulties found in the detection of ternary relationships. The aim of this work is to develop those tools and techniques that allow end-users to correctly design databases and it takes part of a wider project PANDORA, which objective is to define methods and techniques for database development implemented in a CASE tool.

Next section is focused on the methodology applied in the development of the experiment, then we explain the main characteristics of PANDORA Case Tool and how the experiment results might be applied on it. Finally, some conclusions and future work are presented in order to complete this research.

2. Methodology

The experiment has been developed with the main objective of study the difficulties in modelling relationships and their cardinalities, in particular ternary relationship detecion, and analysing those factors that might affect in such process, as the domain knowledge, the requirement presentation within a text and the occurrence of one atributre in a ternary relationship.

In the next section we are presenting the design of experiments and the most outstanding results.

2.1. *Experiment Design*

The data that is being analysed are the answers to a questionnaire that contains eight requirement specifications (textual descriptions), four of them concern to familiar domains and the remaining concern unfamiliar domains. Within this classification, each specification presents a typical characteristic as an indicator for analysing ternary relationship performance difficulties. The Table 1 shows the typology of the tests used in the experiment.

Table 1 shows the test which include ternary relationships, also there are another two tests without ternary relationships for screening the objectives of the experiment to the

[1] N_ary relationships with N>2.

Table 1
Ternary relationship classification

	One sentence / Different sentence	Without intersection data	With intersection data
Familiar	Test 1 (Disco Entrepise) Test 8 (Environmental Department)		Test 5 (Telephone company)
Unfamiliar	Test 2 (Management of legal documents) Test 4 (Review publishers)		Test 6 (Lawyer's office)

users. It is important to distinguish between the same construct in one or several sentences in order to contrast the meaning of the way in which the specifications are written (it is easier to detect ternary relationships when they came from an unique sentence) (Batra and Zanakis, 1994).

For each test, users were able to design an entity-relationship schema and to answer some questions about knowledge domain (familiar or not) and the problems found in the relationship detection as well as some questions about cardinality constraints. Therefore, the variables considered in the experiment are:

1. Degree of domain familiarity (ranging from 0 to 9, where 0 means unfamiliar and 9 is completely known).
2. Difficulties in ternary relationship modelling in terms of degree and cardinality and why.

Notice that the concept of familiar domain is oriented to users who are involved in the experiment. They were students with a variety of knowledge in some domains of database applications. Users were divided into two homogeneous groups, one formed of novices (with a basic database design knowledge) and the other of practitioners.

A group of seventy six students of Computer Engineering studies took part in the experiment. They were highly motivated to participate in the experiment as they got a better qualification in the course of Database Design.

2.2. *Results*

The aim of this experimental study was to identify the difficulties in ternary relationship modelling. The results are dependent on the requirement analysis and the universe of the discourse of each test.

More than half of the individuals were not able to detect ternary relationships (Fig. 1), and those who detect them, do it within a familiar domain in a 60% of the cases (Fig. 2).

As we may see in the results of the Table 2, in which appears the percentage of students who have detected the ternary relationship, the presence of an attribute in the relationship doesn't contribute to detection of it. Nevertheless, if the requirements are exposed in different sentences, the attribute clarifies the design.

It is important to point out that only 2% of the average of students which have modelled the ternary relationship have represented the associated cardinalities properly.

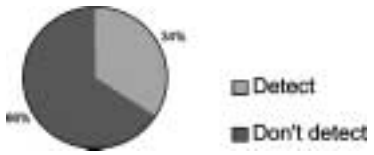

Fig. 1. Students average in the ternary relationship detection.

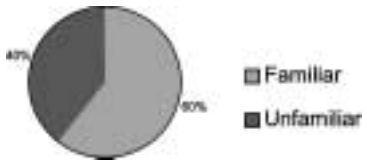

Fig. 2. Students average in the ternary relationship detection (it depend of domain).

Table 2
Percentage of students who have detected the ternary relationship

	One sentence / Different sentence	Without intersection data	With intersection data
Familiar	Test 1 (Disco Entrepise): 36% Test 8 (Environmental Department): 75%		Test 5 (Telephone company): 12%
Unfamiliar	Test 2 (Management of legal documents): 53% Test 4 (Review publishers): 27%		Test 6 (Lawyer's office): 1%

Conceptual modelling is an abstraction process which depends on the ability of the designer, the domain knowledge and the requirement exposition. In particular, there are various notations for cardinality constraints in the ternary relationship representation (Cuadra *et al.*, 2002; Cuadra *et al.*, 2003) so they are more difficult to model.

Additionaly, CASE tools usually implement binary models. So they don't take into account that construct and therefore, novice designers should be helped to model the universe of the discourse through an automatizated tool.

3. PANDORA Project

Database Design Methodologies (Teorey *et al.*, 1986) generally use the Entity Relationship (ER) model (Chen, 1976) as conceptual data model. Next, methodological steps such as logical design, normalisation process and physical design, vary from some methodologies to others; for instance, in some methodologies it is supposed that if a good conceptual design is achieved, the normalisation phase is not necessary (Elmasri and Navathe, 2000; Silberschatz *et al.*, 2001).

Commercial CASE tools for database development do not usually cover database design phases with real EER schemata, and they do not incorporate capabilities for refinement and validation processes. Even, in most cases, they manage hybrid models (merging aspects from EER and Relational models or using a subset of EER graphical notation for representing relational schemata) and sometimes these models are too close to physical aspects.

3.1. *PANDORA Architecture*

PANDORA (CASE Platform for Database development and learning via Internet. Spanish research project CICYT (TIC99-0215)) is a research project devoted to develop a

CASE platform for database learning, design and implementation. It is composed of a set of modules (Fig. 3) that can be independently used or in a methodological framework. In a first level, three layers are identified: conceptual modelling, design and automatic code generation subsystems. Over these layers, there is a learning support subsystem that provides an intelligent tutor for methodological assistance through the use of different tools as well as a Web-based learning component.

The core of PANDORA platform is the *Repository* (metabase) that keeps all the resources and that supports the storage of Extended Entity-Relationship as well as relational schemata, SQL scripts, triggers and so forth. The design of the repository was decomposed in two metamodels, one for storing EER schemata and the other for storing relational schemata. Both of them describe all the constructs that they support. Moreover, this separation clearly distinguishes the two fundamental phases of the database development: conceptual and logical design. Current CASE tools lack this important distinction. Following, a brief description of PANDORA components along with their main contributions are given.

Conceptual Modelling Subsystem is composed of two modules: the EER Modelling and the Natural Language Analysis modules. The former is used by designers in drawing and verifying conceptual schemata; it also allows to store and retrieve conceptual schemata. The Natural Language Analysis module provides some facilities to interpret a descriptive text in order to get proposals of EER schemata according to requirements appearing in the text, (Martinez and Garcia-Serrano, 2000). Moreover, this module also supports an interactive process for identifying and validating binary relationship cardinalities in the conceptual modelling phase. This component profits from natural language processing techniques, first-order logic and some modelling heuristics. Cardinality constraints, especially in higher order relationships, are difficult to understand and model by students, and some validation methods are required. What we propose is an approach that

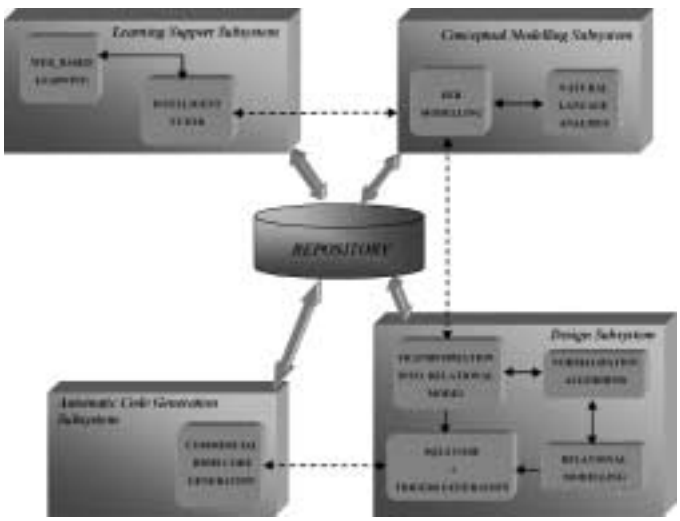

Fig. 3. PANDORA platform.

combines syntax (grammatical categories, word collocations, etc.), semantics (meanings of words, phrases and sentences) as well as first order logic to extract cardinality constraints and validate them with the user.

Design Subsystem includes four modules: Transformation into Relational Model, Relational Modelling, Normalization Algorithms and SQL-3 code (Melton and Simon, 2002) plus Triggers Generation modules. In order to achieve a transformation of EER schemata into the Relational model without loss of semantics, an exhaustive analysis of translating the different EER constructs has been performed. The aim is to develop databases that keep all the integrity constraints an that force their verification regardless of which program accesses the database.

In this subsystem, there are two main contributions: the first one is related to the repository, covering all elements proposed in SQL-3 (Melton and Simon, 2002), including the relational model constraints such as assertions, checks, primary keys, alternate keys and foreign key constraints. Furthermore, inherent constraints are validated by triggers and checks. The second contribution concerns to the relational model transformation, converting the EER constructs into constructors of relational model preserving their associated semantics. A correct transformation of conceptual schemata and their associated constraints is necessary in order to preserve their intended meaning. Although relational model is insufficient for reflecting the complete semantics that can be presented in a conceptual schema, it can be enhanced with specific elements that are used to preserve the original semantics, such as active capabilities (triggers).

In the *Automatic Code Generation Subsystem*, the Commercial DBMS Code Generation module transforms the standard logical schema into an specific logical schema, taking into account the DBMS's characteristics and resolving the relational model's constraints.

Finally, the *Learning Support Subsystem* gives a coherent unification to the CASE environment from two perspectives (Castro *et al.*, 2002). Firstly, the Intelligent Tutor module plays the role of a methodological assistant for guiding the designer during the database development process (through the different phases) and providing support in the design alternatives; the methodology for database development incorporated in PANDORA tool is explained in (De Miguel *et al.*, 1999). Secondly, the Intelligent Tutor incorporates some teaching and training strategies of database design concepts that can be used via Web. For this purpose, a set of didactic units together with a set of exercises, have been designed (Iglesias *et al.*, 2002).

In order to perform this pedagogic component we collaborate with people from the New Information Technologies and Communication Program (PNTIC) belonging to the Spanish Ministry of Education, Culture and Sports that has a wide experience in designing and implementing Web courses for distance learning. They help us to define the didactic units and also they will provide us a platform to test the Learning Support subsystem with a national coverage.

3.2. *Methodological Assistant of Cardinality Constraints*

In this section we are describing the assistance that the user may request to PANDORA for finding or validating cardinality constraints through the HACIB (Binary relationship

cardinalities help) component. This component is called by the Intelligent Tutor to give support in database design learning to teach the cardinality constraint concept (Fig. 4 shows a snapshot of the interface).

The interpreter for cardinality constraints is implemented in Prolog using the Definite Clause Grammars (DCG) formalism.

An analysis of a wide corpus of Spanish short descriptive texts describing several UoDs has allowed us to identify the linguistic elements (determiners, adverbs, adjectives and others) that help to obtain entity and relationship types as well as cardinality constraints. The idea is to translate each sentences into a logical formula, taking into account some rules of scope and precedence of quantifiers, in order to validate semantics with the user.

The process has been implemented in PROLOG[2] language and it is divided into three steps, the last one divided itself into two tasks.

Briefly, first step is in charge of processing the Natural Language (NL) sentence to obtain a logic formula, second step gives the user the possibility of entering data in order to validate the semantic; this validation is carried out in the third step with two different behaviours.

Step 1: Analyse the sentence using a DCG grammar, which rules have syntactic, semantic and pragmatic features.

Step 2: The user is given the choice of introducing domain values corresponding to the semantic predicates that appear in the analysed sentence.

Step 3: Validate the semantic formula F with the domain values instantiated by the user in Step 2.

Step 3a: If the formula F is true, the system informs users about the cardinalities assigned to the relationship obtained using mapping rules.

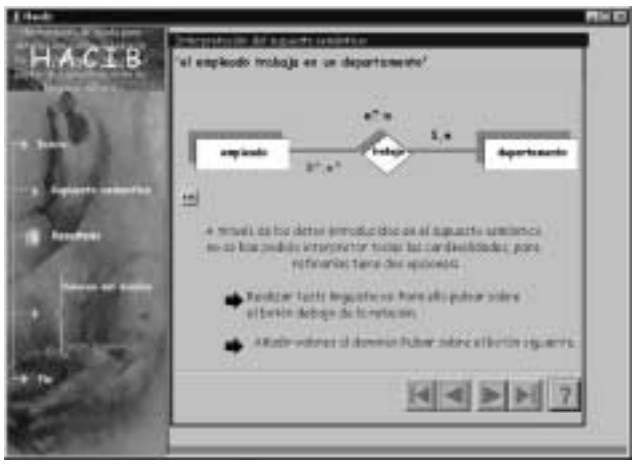

Fig. 4. Identifying and validating cardinality constraints.

[2]AMZI! Prolog v 4.1

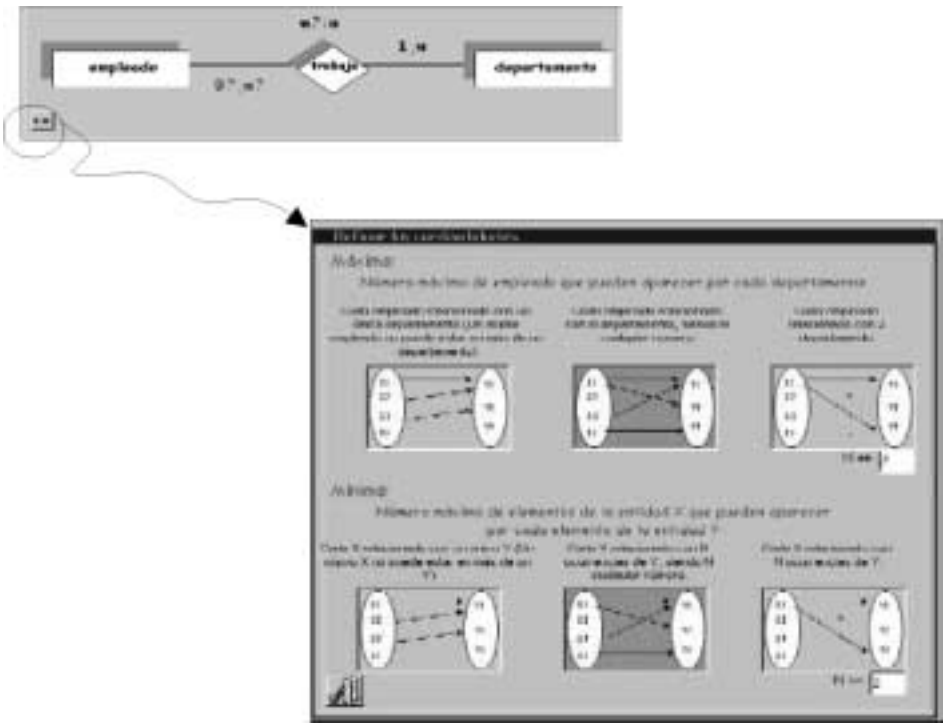

Fig. 5. Linguistic test in HACIB.

Step 3b: If the formula F is false, there are two possibilities: to select a new sentence interpretation or to trigger some linguistic tests.

It is important to stress that linguistic test for detecting ambiguous cardinalities (case of F false) has been achieved by Venn Diagrams (Fig. 5), as it is proposed by Elmasri and Navathe (2000) and also with the semantic interpretation of each alternative.

HACIB interprets and validates natural language sentences about binary relationships, but at present, grammar extensions have been implemented in order to incorporate ternary relationship analysis. Since sentences about that construct hold more ambiguity than binary relationships, a dialog with the user must be carried out to obtain respective cardinalities.

4. Conclusions and Future Work

This paper presents our research in progress on identifying ternary relationships and the control of cardinality constraints. At present, we have detected that novices and practitioners have similar behaviour in terms of detecting ternary relationships and their associated cardinalities. Most frequently problems are focused in the domain knowledge and the way in which the constructs are transcribed in the texts.

After a first review of the student tests, some preliminary results can be outlined:

1. If the sentences concerning the ternary relationship are together in the same paragraph, it is easier for the student to detect it. However, if the sentences describing

the ternary relationship are distributed along the textual specification, students tend to represent it as a combination of binary relationships.

2. If the ternary relationship contains one attribute (data intersection), usually the detection of this construct is easier than if there were no attribute in the relationship.
3. If there is some implicit knowledge (common sense knowledge) in a familiar domain, students found some difficulties in detecting ternary relationships.

At this moment HACIB only works with binary relationships. Currently, we are developing a new version of HACIB to collect ternary relationships and the posibility to generate a natural language interpretation from the schemata, due to the difficulties to detect n-ary relationships within natural langauge test sentences.

Also some experiments must be done with HACIB and our students, in order to check the level of methodological assistance that it provides.

References

Batra, D., and S.R. Antony (1994). Novice errors in conceptual database design. *European Journal of Information Systems*, **3** (1), 57–69.

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

Batra, D., and H. Zanakis (1994). A conceptual database design approach based on rules and heuristics. *European P. Journal of Information Systems*, **3** (3), 228–239.

Bock, D., and T. Ryan (1996). Modelling ternary relationships. *Journal of Computer Information Systems*, 60–65.

Castro, E., *et al.* (2002). Integrating intelligent methodological and tutoring assistance in a CASE platform: the PANDORA experience. In *Proceedings of the Informing Science & IT Education Conference*. Cork, Irland.

Chen, P. (1976). The entity-relationship model – towards an unified view of data. *ACM Transactions on Database Systems*, **1** (1), 9–36.

Cuadra, D., *et al.* (2002). *Preserving Relationship Cardinality Constraints in Relational Schemata*. Database Integrity: Challenges and Solutions, Idea Group Publishing.

Cuadra, D., *et al.* (2003). *Dealing with Relationship Cardinality Constraints in Relational Database Design*. Effective databases for text & document management. Idea Group Publishing.

De Miguel, A., *et al.* (1999). *Diseño de Base de Datos Relacionales*. RAMA.

Elmasri, R., and S.B. Navathe (2000). *Fundamentals of Database Systems*. Addison-Wesley.

Goldstein, R., and V. Storey (1990). Some findings on the intuitiveness of entity-relationship constructs. In *Entity Relationship Approach to Database Design and Querying*. Elsevier Science Publishers, Holland, pp. 9–23.

Hitchman, S. (1995). Practitioners perceptions on the use of some semantic concepts in the entity-relationship model. *European Journal of Information Systems*, **4**, 31–40.

Iglesias, A., *et al.* (2001). Learning to teach database design by trial and error. In *4th International Conference on Enterprise Information Systems*. Ciudad Real, Spain, pp. 500–505.

Martinez, P., *et al.* (1999). Profundizando en la semántica de las cardinalidades en el modelo E/R extendido. In *IV Jornadas de Ingeniería del Software y Bases de Datos*. Spain, pp. 53–54.

Martinez, P., and A. García-Serrano (2000). On the automatization of database conceptual modelling through linguistic engineering. In *Natural Language Processing and Information Systems* (NLDB). Revised papers, LNCS 1959, pp. 276–287.

Martinez, P., *et al.* (2000). Data Conceptual Modelling through natural language: Identification and validation of relationship cardinalities. In Idea Group Publishing (Ed.), *Challenges of Information Technology Management in the 21st Century*. ISBN, Anchorage, pp. 500–504.

Melton, J., and A.R. Simon (2002). *SQL: 1999. Understanding Relational Language Components*. Morgan Kaufmann Publishers.

Siau, K., *et al.* (1995). A psycological study on the use of the relationship concept. Some preliminary findings. In *Conference on Advanced Information Systems Engineering* (CAISE). Finland, pp. 341–354.

Silberschatz, A., *et al.* (2001). *Database Design Concepts*. 4th ed. McGraw-Hill.

Teorey, T.J., *et al.* (1986). A logical design methodology for relational databases using the extended entity-relationship model. *ACM Computing Survey*, **18** (2).

E. Castro received the MSc in mathematics from Universidad Complutense of Madrid in 1995. Since 1998 she works as assistant lecturer at the Advanced Databases Group in the Computer Science Department of Universidad Carlos III of Madrid. She is currently teaching relational databases. Her research interests include database conceptual and logical modelling, advanced database CASE environments and information engineering.

D. Cuadra received the MSc in mathematics from Universidad Complutense of Madrid in 1995. Since 1997 she works as assistant lecturer at the Advanced Databases Group in the Computer Science Department of Universidad Carlos III of Madrid. She is currently teaching database design and relational databases. Her research interests include database conceptual and logical modelling and advanced database CASE environments.

P. Martínez got a degree in computer science from Universidad Politécnica of Madrid in 1992. Since 1992, she has been working at the Advanced Databases Group in the Computer Science Department at Universidad Carlos III of Madrid. In 1998 she obtained the PhD degree in computer science from Universidad Politécnica of Madrid. She is currently teaching database design, advanced databases in the Computer Science Department at the Universidad Carlos III de Madrid. She has been working in several European and National research projects about natural language processing, advanced database technologies, knowledge-based systems and software engineering.

Empirinis požiūris į trinarių sąryšių naudojimą duomenų bazių koncepciniame modeliavime

Elena CASTRO, Dolores CUADRA, Paloma MARTÍNEZ

Informacinių sistemų projektavimas apima tokius svarbius aspektus kaip koncepcinis bei loginis modeliavimas, nagrinėjimo srities žinių supratimas. Išplėsto esybių sąryšių (Extended Entity Relationship – EER) modelio taikymas yra parankus būdas supažindinti su reikalavimais informacinėms sistemoms: šis modelis pasižymi pakankamai aukštu abstrakcijos lygiu, tad jį pritaikius galima surinkti visą diskurso aibės semantiką. Ir nors vartotojui tai yra artima bei paranku, sunkumų kyla tada, kai panaudojami tokie komponentai, kaip aukštesnio lygio sąryšiai. Be to, ir nagrinėjimos srities žinios gali įtakoti modeliavimo procesą. Atsižvelgus į tam tikrus EER modelio taikyme kilusius sunkumus buvo atlikti eksperimentai, kuriais siekta išsiaiškinti problemas kylančias dar projektavimo stadijoje. Šiame straipsnyje supažindinama su atliktu eksperimentu (bei gautais keliais svarbiais rezultatais), kuriame dalyvavo duomenų bazių projektavimo srities dviejų skirtingų lygių studentai. Šiuo eksperimentu buvo siekiama atskleisti kai kuriuos esamus sunkumus, kurie kyla tiek pradedantiesiems studentams, tiek specialistams kuriant duomenų bazes bei naudojant trinarius sąryšius. Aptariamos ir kai kurios specialios temos nustatant trinarius sąryšius, tokios kaip nagrinėjimo srities svarba, reikalavimų pateikimo būdas bei sankirtos duomenys. Siekiant vadovauti ir padėti vartotojams, sprendžiantiems projektavimo uždavinius, šie rezultatai yra panaudoti ir PANDORA instrumentinės programinės priemonės moksliniame tiriamajame projekte, kuriame yra apibrėžiami duomenų bazių kūrimui skirti metodai ir priemonės bei realizuojama metodinė pagalbinė priemonė, naudojanti koncepcinio modeliavimo bei mokymosi pagalbos posistemius.