Comparative Study of Four Data Modeling Approaches

J.H. ter Bekke

Department of Information Systems
Delft University of Technology
P.O. Box 356, 2600 AJ Delft, The Netherlands
E-mail: j.h.terbekke@is.twi.tudelft.nl
Telephone: +31 - 15 - 2784402
Fax: +31 - 15 - 2786632

Abstract

The regular national examination on "Analysis, modeling and management of data" in The Netherlands has been used for an analysis of four modern data modeling approaches, namely: relational, semantic, entity-relationship and binary. Comparison of these approaches was possible because there were enough candidates with the same educational training in data modeling. The analyzed examination session gave candidates for the first time in many years the possibility to choose their own favourite data modeling approach for the design case in the examination work.

This paper starts with an overview of the main concepts of the four data modeling approaches. The description and standard solutions of the design case are given together with illustrations of some major characteristics of the four approaches. Analyses of the scores indicate among others that better candidates choose for the semantic approach. Their overall scores were therefore also higher.

1 Background

The EXIN Foundation organizes national computer science examinations in The Netherlands. The examination on "Analysis, modeling and management of data" [2] is held twice a year for hundreds of candidates from several educational institutes in The Netherlands. Subjects are four modern data modeling approaches: relational, semantic, entity-relationship (here EAR) and binary. Candidates must have knowledge of basic concepts, data modeling in all four data modeling approaches and must be capable to compare modeling results.

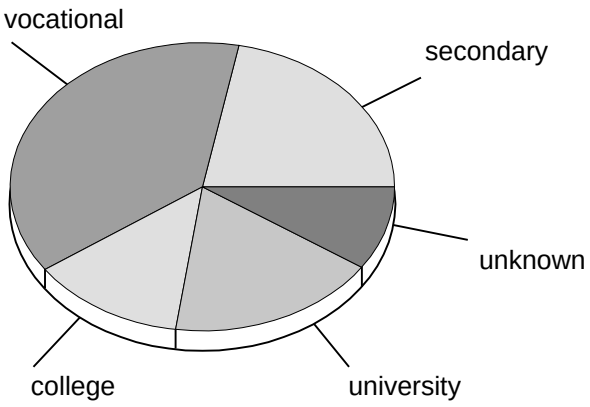
The examination session consists generally of a part with multiple choice questions regarding all subjects (containing questions homogeneous spread over all four modeling approaches) and a large case description which must result in a design according one of the four modeling approaches. The whole examination session takes 3 hours of which 75 minutes for design. Generally the modeling approach is prescribed. The analyzed examination session however gave candidates for the first time the possibility to choose their own favourite modeling approach for the design case. This offered the possibility to compare results of the modeling approaches. In cooperation with EXIN Foundation some analyses have been carried out on the individual examination scores. This resulted in some remarkable conclusions regarding the modeling approaches.

2 Purpose of the study

There are several data modeling approaches. The approach used in practice does not have to be the one producing the best results. It is often determined by the standards used in an organisation. This can make things rigid. Modern approaches, leading to better results, are therefore not recognized. This situation will not change easily, moreover because often comparison is not feasible. For example, it is not possible to solve the same case according different modeling approaches by the same design team. In such a situation earlier designs will influence later designs.

The purpose of this study is to discover relationships between the modeling approach in use and the resulting quality. The examination session in question had over two hundred candidates with the same educational training in data modeling. This session could offer therefore a possibility to make some statements about the four approaches.

The population consisted of candidates with different general educational background varying from secondary to university level. The percentage of candidates who qualified was between 52 and 70 percent (see figure 1). The course demands 180 hours of study; i.e. 35 hours per data model.



background	number	successful
secondary	46	52 %
vocational	79	56 %
college	27	67 %
university	37	70 %
unknown	20	55 %

Figure 1: Educational background and percentages of successful candidates

3 Overview of the four approaches

The design case of the examination work requires knowledge and application of concepts of four data modeling approaches. For the examination in question only a limited collection of concepts could be used, because the solution had to be expressible in all four data modeling approaches. These concepts can be summarized by the semantic concepts of type and aggregation. Below is an overview of the main concepts in all four approaches using the same simple library example.

The relational data model uses the concept of relation as its data structuring concept [1]. A relation can be conceived as a two dimensional table (considered as a set) containing only atomic values. The definitions below are some examples:

reservations (**book-id, member-id**)
members (**member-id**, name, address, postal code, city)
books (**book-id**, title, author, publisher)

Relation ’reservations’ has two attributes (i.e. table columns). The number of rows in a relation depends of the number of data stored in the table. Each row has a unique identification consisting of one or more table column values. This uniqueness constraint is expressed in the so-called primary key of the relation (here in bold). The coherence (in particular 1:n relationships) between relations is expressed by foreign keys (here underlined); they contain references to other relations. An example is attribute book-id from the ’reservations’ relation containing references to the ’books’ relation. This property is known as referential integrity.

In the semantic data model the concept of type plays a dominant role [5]. A type is defined as the collection (aggregation) of a definite number of properties into a unit. These properties themselves are also types. The library example above is specified as follows:

type reservation = book, member.
type member = name, address, postal code, city.
type book = title, author, publisher.

Aggregation can also be represented in an abstraction hierarchy. The foregoing example is represented in figure 2. The aggregation is always placed above its properties. Base types (as for example: name, address, etc.) are not represented in the graphical representation. Note that referential integrity is inherent in type definitions and abstraction hierarchies.

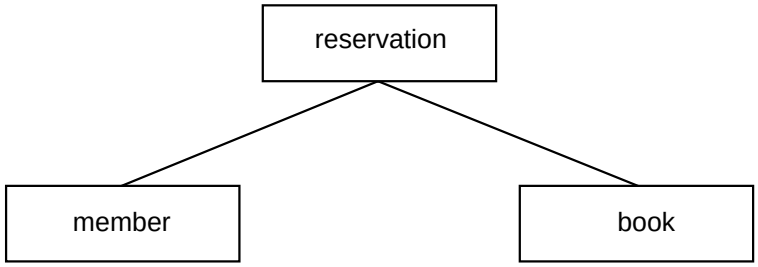


Figure 2: Abstraction hierarchy

The EAR data model [3] is based on the concepts of entity type, attribute type and relationship type. The concept of entity type is the structuring concept; it contains a number of elementary attributes. Next to this, the concept of relationship type is used. Relationships are mostly information bearing. The previous simple library example can be expressed in an EAR diagram as follows (see figure 3).

Figure 3: EAR diagram

The parameters c1 and c2 in figure 3 denote (minimum, maximum) cardinalities. Usually these values are 0, 1 or n. The complete EAR model consists of relationship type and entity type definitions including identification. This example has two entity types: ’member’ (with

identification the attribute member-id), 'book' (with identification the attribute book-id) and one relationship type 'reservation'.

The binary model [4] contains many concepts. Essential is the distinction between lexical and non-lexical object types. Lexical object types (denoted by a dotted circle) are object types that can be used as names for or references to other object types; non-lexical object types (denoted by a closed circle) are named object types or object types referred by lexical object types. A bridge type is a relationship between a lexical and a non-lexical object type. The relationship between two non-lexical object types is called an idea type. Graphical constraints can be imposed on a information structure diagram, among them: the uniqueness constraint (u) and the totality constraint (v). Each binary model must be referrable, that is: it must always be possible to refer uniquely to a non-lexical object type. The foregoing simple library example is represented in an information structure diagram in figure 4. This graphical representation is more complex than the others.

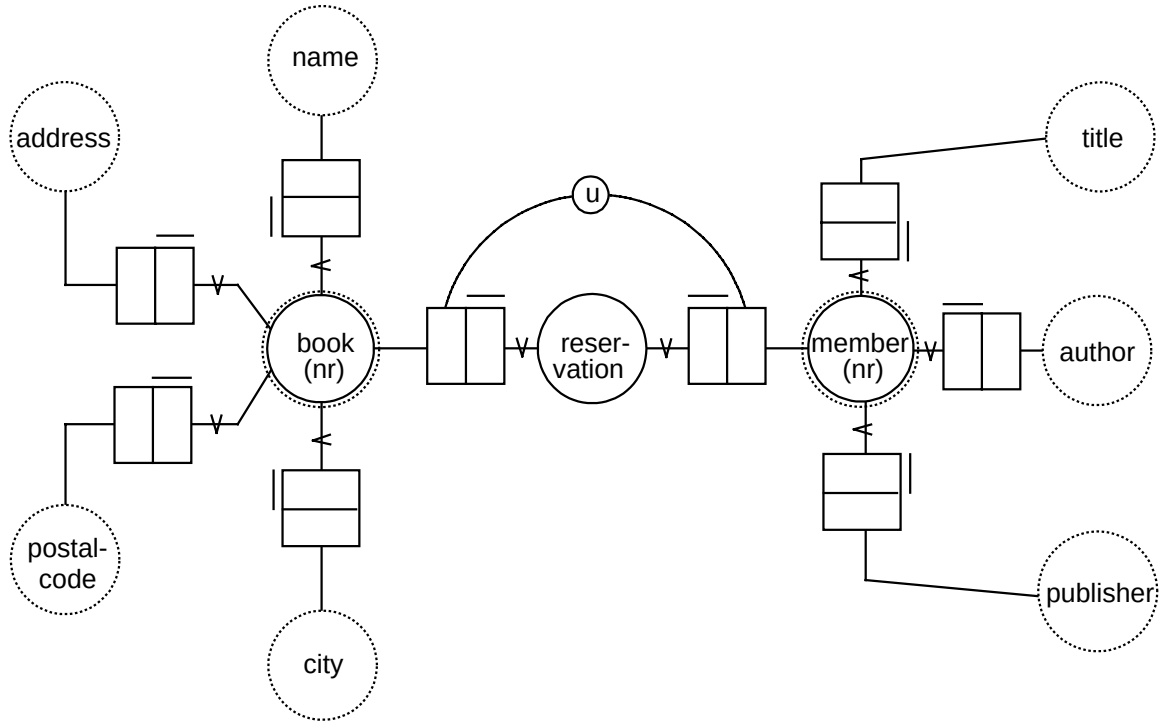


Figure 4: Information structure diagram

4 The design case

Limitations

Certain limitations are imposed on the case description. The description must be such that a solution can be developed in all four approaches. This implies that certain aspects occurring in semantic models (as specializations and generalizations) cannot be included in the case; these aspects occurred only in the multiple choice part of the examination work. Another requirement was the homogeneity of the evaluation. It implied an evaluation in which only structural aspects could play a role.

Case description

A university has in each faculty a unit with organisational and registrational responsibilities concerning practical works. Students can participate during certain non-overlapping periods in a practicum. An identification, a description and the responsible lecturer are registered for each practicum. For each lecturer are registered: identification, name and telephone number. It is registered how many exercises are required per practicum. The standard exercises, uniquely determined for a practicum, are also registered. For each practicum period are registered: starting date, closing date, a certain weekday and the opening and closing hour of the laboratory on that weekday. The university has several faculties, each with a unique name, an address and a city. Some faculties have more than one field of study. Each field of study belongs to one faculty and is characterized by a name and a description.

Students are enrolled in one field of study under identification and session (i.e. class). Their name, address, city and date of birth are also registered. Sometimes a practicum is required for more than one field of study. A practicum can be registered in different sessions and different fields of study. Based on the field of study in which a student is enrolled, it must be possible to determine whether or not a student may participate in a practicum. Each participant in a practicum receives an evaluation for the completed participation. For each standard exercise are registered: the estimated time needed to complete the exercise, a description of the exercise (unique within the practicum), a category (for example: optics, thermodynamics, electronics) and in which practicum it occurs. There are standard exercises for all categories. It must be possible to determine the number of standard exercises per practicum. Enrolment of a student for a practicum period implies the provision of certain individual works. A work is related to a standard exercise. The number of works correspond with the number of required number per practicum. Each completed work must result in a report by the student. This report is evaluated by the lecturer and results in a mark.

5 Four solutions

The foregoing introduction to the four approaches has made it evident that the binary solution will be very laborious. The candidates were also warned for this. The standard solutions according all four approaches are given below.

Relational model

Required in the solution are relations including the **primary** and foreign keys:

lecturer	(**lecturer-id**, name, telephone-number)
practicum	(**practicum-id**, description, lecturer-id, required-number)
faculty	(**faculty-name**, address, city)
field	(**field-name**, description, faculty-name)
student	(**student-id**, session, name, address, city, day-of-birth, field-name)
period	(**practicum-id**, **starting-date**, closing-date, weekday, opening-hour, closing-hour)
subject	(**practicum-id**, **field-name**, session)
participation	(**student-id**, **practicum-id**, **starting-date**, evaluation)
exercise	(**practicum-id**, **description**, estimated-time, category)
work	(**description**, **practicum-id**, **starting-date**, **student-id**, mark)

Semantic model

Required are type definitions and corresponding abstraction hierarchy (figure 5):

type lecturer = name, telephone number.
type practicum = description, required_number, lecturer.
type faculty = name, address, city.
type field = description, faculty.
type student = name, address, city, birth_date, field, session.
type period = practicum, starting_date, closing_date, weekday, opening_hour, closing_hour.
type subject = practicum, field, session.
type participation = period, student, evaluation.
type exercise = description, estimated_time, category.
type work = exercise, participation, mark.

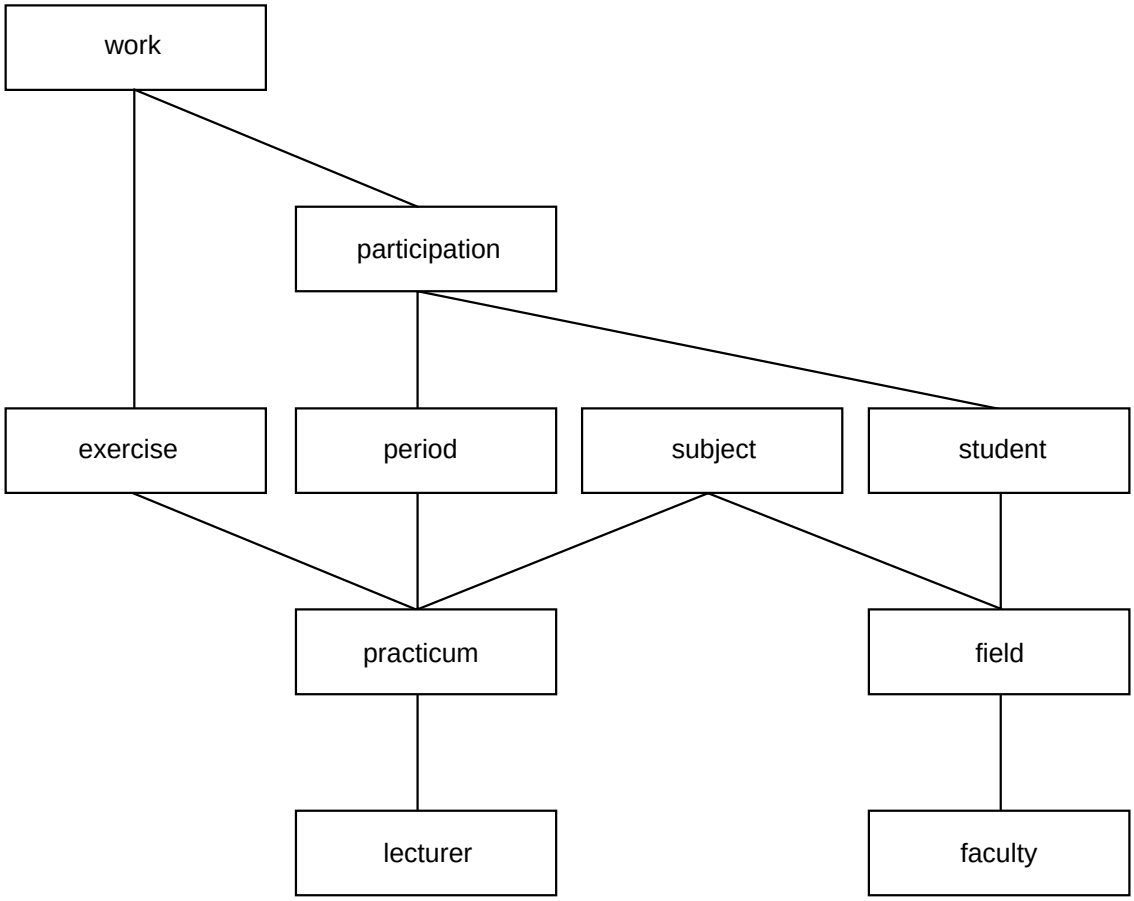


Figure 5: Abstraction hierarchy for the practicum organisation

EAR model

Required are the graphical representation (including entity types and relationship types) and a separate collection of entity types including identification. The solution in figure 6 contains '-' in case both cardinalities 0 and 1 are allowed.

Entity type:	*Identification:*
lecturer	lecturer-id
practicum	practicum-id
faculty	faculty-name
field	field-name
student	student-id
period	practicum-id, starting-date
subject	practicum-id, field-name
participation	student-id, practicum-id, starting-date
exercise	practicum-id, description
work	description, practicum-id, starting-date, student-id

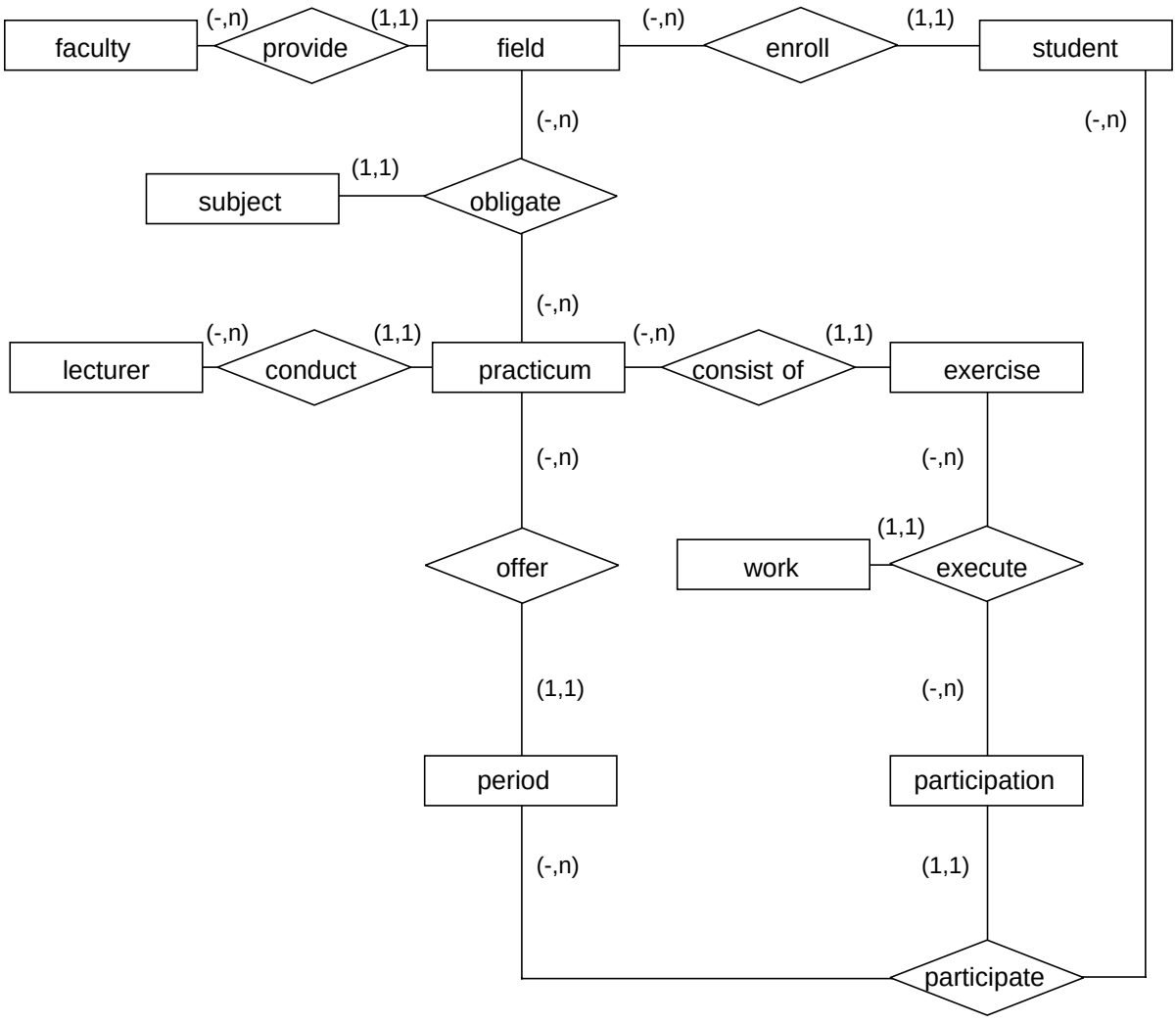


Figure 6: EAR diagram for the practicum organisation

Binary model

Required is a referrable information structure diagram (figure 7).

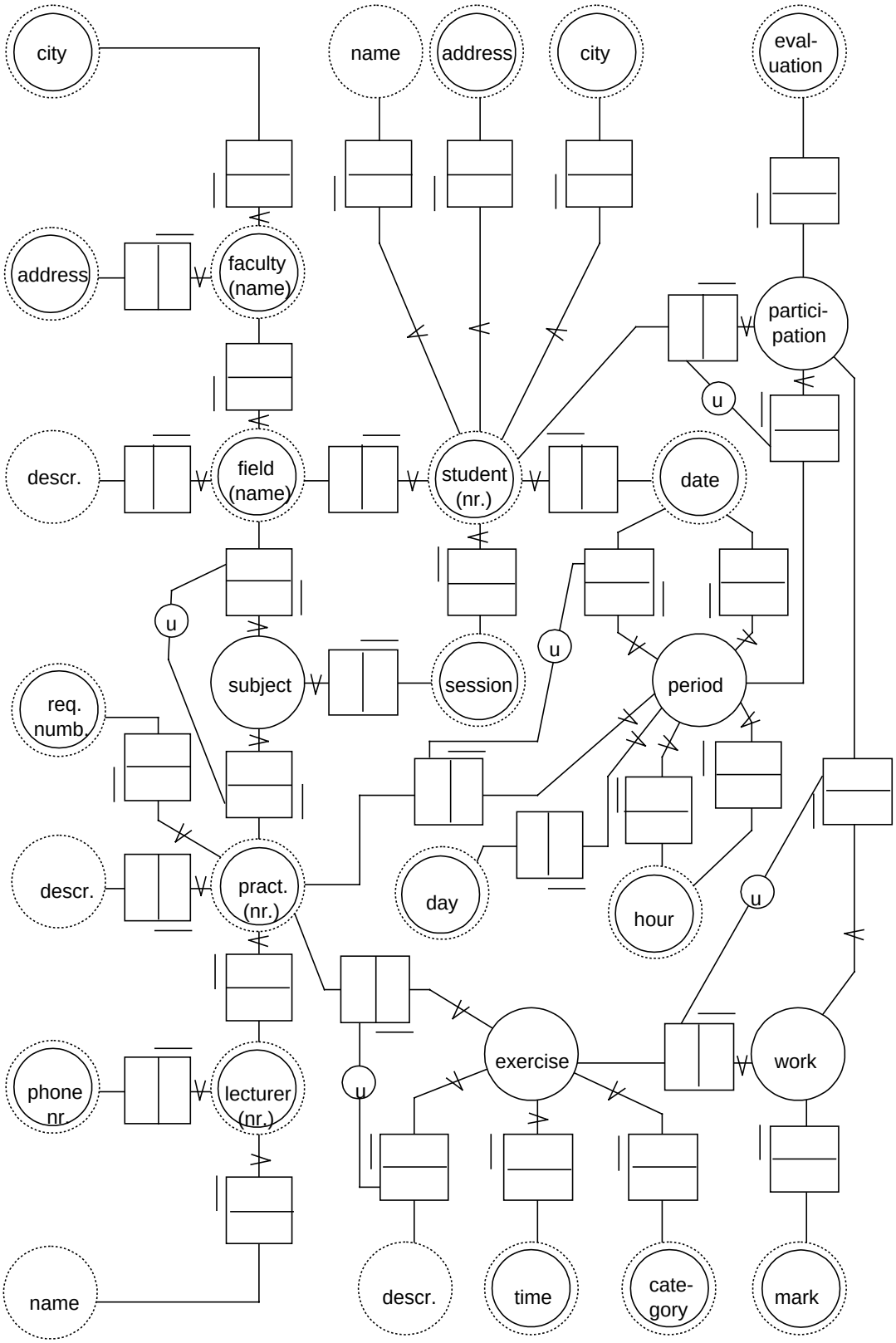


Figure 7: Information structure diagram for the practicum organisation

6 Some characteristics of the approaches

Foregoing standard solutions illustrated a number of characteristics of the four approaches. They occur generally in designs according these approaches. A designer who uses such an approach must be aware of these characteristics.

Relational model
Application of the relational concepts leads often in the use of primary keys consisting of several attributes. In the relational solution this occurs for example in the definitions of the relations 'exercise' and 'work'. These large keys have in practice several unwelcome effects:

- Attributes who contribute to a primary key may not be absent and may not be changed. For example the attribute 'starting date' of 'practicum' may not be changed. It is evident that this would disconnect 'practicum' from 'participation'.
- Specification of desired relationships by join operations in relational languages will lead to serious performance degradations. This degradation is visible in the enforcement of referential integrity, but also in the use in data manipulation or query commands.
- The probability of overlapping foreign keys is increased by the requirements of normalization procedures (such as BCNF [1]). This will also result in different meanings for the same attributes of a relation. The aspect can be discovered in relation 'work' in which 'practicum-id' contributes to the reference from 'practicum' via 'participation' and 'period' as well as the reference from the same 'practicum' via 'exercise'. In this case it does not result in semantic problems because both refer to the same 'practicum'. Generally this is not the case.

Semantic model
The semantic model introduces for each type a simple identification. This results in unambiguous references. The problem of references via different semantic relationships is directly visible in the abstraction hierarchy. In this case the following additional static constraint is required:

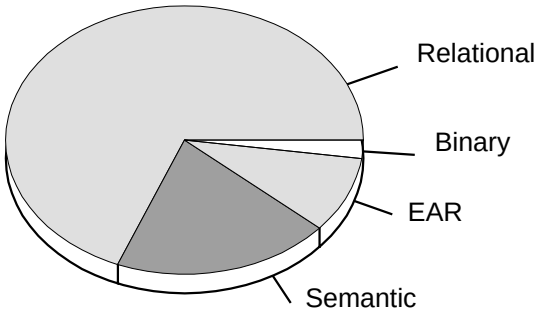
assert work *its* allowed (*true*) =
    exercise *its* practicum = participation *its* period *its* practicum.

EAR model
Relationships in an EAR model are for (1, n) cardinalities often information bearing. Also (1, 1) cardinalities are often required in a design. These relationships are in a certain way superfluous and are caused by the concepts on which the modeling approach is based. For instance the entity type 'subject' is only required because of the attribute 'session'. A static constraint (as in the semantic solution) cannot be expressed.

Binary model
The information structure diagram shows that the distinction between lexical object types and non-lexical object types is often artificial. Besides that, a binary solution is hardly verifiable. Despite of all efforts to specify all constraints graphically, it is not possible to specify static constraints as in the semantic solution.

7 Analyses of the scores

The first analysis concerns the choice of the candidates. It is obvious that the relational model has a high score here because of the many relational implementations in practice. An overview of this choice is given in figure 8.



MODEL	CHOICE	
	overall	top 10 %
Relational	69 %	52 %
Semantic	20 %	48 %
EAR	9 %	-
Binary	2 %	-

Figure 8: Choice of the candidates

It is surprising that a rather high percentage has chosen for the semantic model (even 48 % of the top ten percent in overall score). The small number of candidates for the binary model is also caused by the given warning.

Interesting are also the scores obtained for the design case. The percentage of candidates who succeeded with a particular choice is compared with the percentage of the same candidates who succeeded in the multiple choice part of the examination (see figure 9).

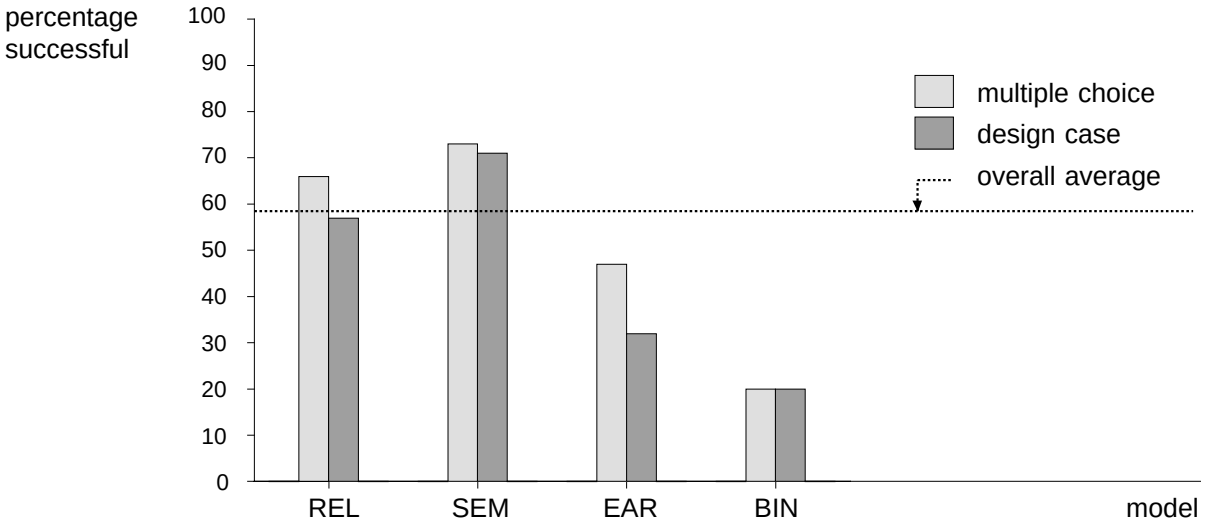


Figure 9: Percentages of successful candidates

From figure 9 can be concluded that the better candidate chooses the semantic model and that overall score is only a little bit influenced by this choice. For other data models a remarkable

degradation can be discovered. Weaker candidates like to use a data model with many prescriptions like the EAR model and the binary model. However, the overall score has a negative influence on this. The scores for the binary model are not very useful because of the small population (2 %) and the given warning.

The next analysis has to do with the partition of scores in a scale from 2 (low) to 9 (high) for the design case and for the complete examination, see figure 10. The population is divided per data model. A student is considered to be successful with a score of 6 or higher.

Again, also from figure 10 can be concluded that the better candidates selected the semantic model. This group was also able to improve their overall score by this choice. This appears from the distribution for the design case: only a few who did not succeed, and a concentration near 7 and 8 and the maximum of 27 % at 9. The semantic model is therefore educationally good.

In case of the relational model the influence on the final score is higher. This can be concluded from the flat distribution for the design case. However, the influence of the EAR model and the binary model is much bigger. In fact the overall score is negatively influenced by these choices. The number of candidates who failed is remarkably bigger than the number of candidates who succeeded.

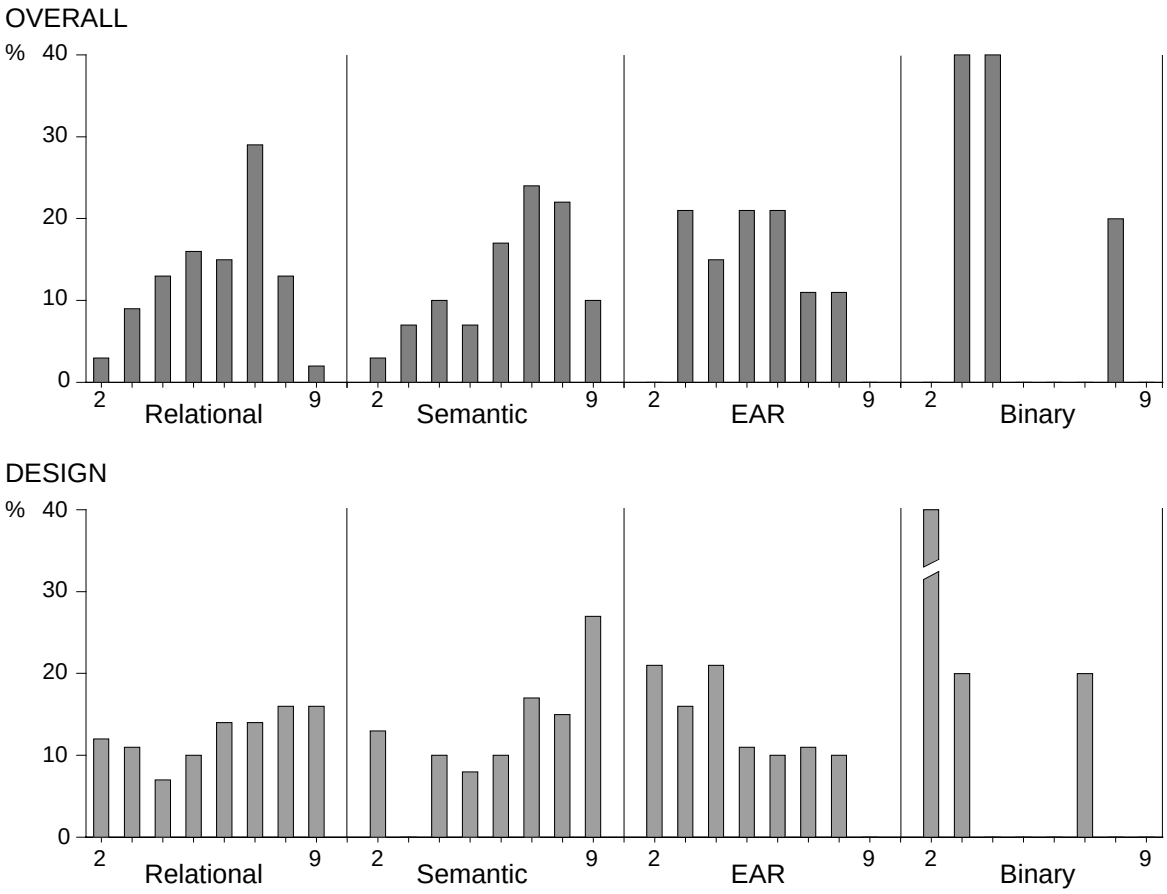


Figure 10: Partition of scores for complete examination and design case

8 Limitations

The described analysis is related to one examination session in which more than two hundred candidates participated. These candidates came from several educational institutes. Unfortunately, because of the small number of candidates who selected the entity-relationship and the binary approach, the analyses regarding these data models are not very useful. To get a better idea of the ’performance’ of these data models an analysis must be continued with examinations in which these models must be used for a design case.

An analysis with results from several examination sessions needs some additional remarks. The complexity of a design often determines the prescribed data model. A design for the semantic model may contain aspects specific for this data model such as the abstractions generalization and specialization and the facility to specify so-called blocks of mutually disjunct specialization types. These specifications are not always possible in other data models.

A case description for the binary model is generally much simpler than the case in this paper because of the number of details that must be specified. The entity-relationship model requires a very strict description in terms of cardinalities. In spite of this, as in the analyzed examination, we see much freedom in the specification of cardinalities. The relational model takes in all these requirements a middle position.

The requirements of the examination on "Analysis, modeling and management of data" seldom allow an examination session as the one reported on. This because a candidate must be tested in knowledge of all four data models. This implies that an analysis such as described in this paper can be done only by exception. A repeat on a regular basis of this kind of examination would lead to a situation in which candidates prepare themselves in only a subset of the four data models.

Acknowledgments

This paper would not be accomplished without the willingness of EXIN Foundation. I owe this organisation therefore many credits.

References

[1] C.J. Date, An introduction to database systems, Vol. I, Addison Wesley, Reading Mass. (1990).

[2] EXIN Foundation, AMBI 88 Examenplan, Kluwer/Examenpublikaties informatica, Maarssen NL, (1993).

[3] ISO/TR9007, Information processing systems - concepts and terminology for the conceptual schema and the information base, ISO (1987).

[4] G.M. Nijssen and T.A. Halpin, Conceptual schema and relational database design, Prentice Hall, Hemel Hempstead UK (1989).

[5] J.H. ter Bekke, Semantic data modeling, Prentice Hall, Hemel Hempstead UK (1993).