



## UWS Academic Portal

### The concept of representation capability of databases

Feng, Junkang

*Published in:*  
Journal of Data Analysis and Information Processing

Accepted/In press: 05/04/2019

*Document Version*  
Peer reviewed version

[Link to publication on the UWS Academic Portal](#)

*Citation for published version (APA):*  
Feng, J. (Accepted/In press). The concept of representation capability of databases. *Journal of Data Analysis and Information Processing*, 1-8.

#### General rights

Copyright and moral rights for the publications made accessible in the UWS Academic Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

#### Take down policy

If you believe that this document breaches copyright please contact [pure@uws.ac.uk](mailto:pure@uws.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.

# The Concept of *Representation Capability* of Databases

**Junkang Feng**

Database Technology Research Group  
School of Computing, Engineering and Physical Sciences  
University of the West of Scotland, UK  
Email: junkang.feng@uws.ac.uk

**Abstract**—The *representation capability* of an information system in general and a database in particular seems an important and yet elusive concept, which is concerned with in our view how a database ever becomes capable of representing real-world objects accurately or otherwise. To explore how to approach and then define this concept, we explore what is meant by that a database connection (i.e., a connection between database constructs such as entities in an Entity-relationship (ER) diagram and relations in a relational schema that are made available by a database) *refers to*, *represents* and *accurately represents* a real-world relation respectively. We also find that that the *information content* of the former includes the latter is a sufficient and necessary condition for the former to be able to accurately represent the latter. All these make the concept of representation capability of a database approachable and definable.

**Index Terms**—Representation Capability, Database modelling, Database theory, Information content

## I. INTRODUCTION

The motivation for this work is to explore what enables and is required for a database to represent real-world objects accurately or otherwise, in other words, how a database becomes capable of representing real-world objects accurately or otherwise and thus the *representation capability* of databases. Gregor's paper [1] in MIS Quarterly says that 'Calls continue for "good theory"' [2] and 'the development of our "own" theory' [3] and presents the nature of theories in information systems. The questions that arise about the bodies of knowledge or theories encompassed in a discipline fall into a number of inter-related classes, and the first one is 'domain questions' [1]. Such questions are concerned with what phenomena are of interest in the discipline, and what the boundaries of the discipline are. We believe that the representation capability of an information system in general and that of a database in particular should be within the boundaries of the

discipline of information systems including databases.

To this end, we explore what is meant by that a database connection (i.e., a connection between database constructs such as *entities* in an ER diagram and relations in a relational schema that are made available by a database) *refers to*, *represents* and *accurately represents* a real-world relation respectively. We also find that that the *information content* of the former includes the latter is a sufficient and necessary condition for the former to be able to accurately represent the latter. All these constitute a seemingly effective means to approach the important and yet elusive concept of the representation capability of databases. To develop our solution, we draw on *semiotics*, *the semantic theory of information* presented by Dretske [4] and *the information channel theory* by Barwise and Seligman [5].

## II. A SEMIOTIC PERSPECTIVE FOR DATABASES

In order to explore how a database construct becomes capable of representing certain real-world objects, we propose an approach that is based on the ideas of semiotics [6], [7]. Semiotics is the study of signs or the general theory of representation [8]. Semiotics has been used to tackle problems in information systems development. For example, Sian and Tian in reference [8] suggest that the graphical notions (or visual signs) of UML are subjected to the principles of signs, and therefore they use semiotics to study the effectiveness of them. We view a database as a collection of signs, and the real-world objects that a database represents are seen as part of properties of signs. Moreover, Stamper [6] points out, 'signs on every level depend for the correct formation of signs on the level below.' Therefore a database can be looked at, at least, on two different levels – syntactic and semantic. The former is concerned with the formal structure of the database, and the latter objects and relationships among them that the

signs (i.e., data) and constructs of the signs signify. A database design problem may be viewed as a mismatch between the two levels.

#### A. Database Connections vs. Real-World Relations

In the context of conceptual database schemata, two types of connections are in question. The connections between data constructs, such as ‘entity’, that are made possible by the topological structure (i.e., a syntactic level formation of signs) of a conceptual database schema or diagram can be termed ‘database connections’ without considering what in the real-world to which they refer. The connections between real-world objects, which is what we want represented by using ‘database connections’, may be called ‘real-world relations’. They are independent of a modelling mechanism such as ER. For example, it might be a real-world fact that employee  $e_1$  belongs to division  $d_1$ , which would be a ‘real-world relation’. If two entity instances, say node  $e_1$  and node  $d_1$ , are connected by an edge in the instance diagram of an ER diagram such as the lower half of Fig. 2, then there is a database connection between them.

A basic task in database design is to construct a sufficient (minimally sufficient if possible) conceptual database diagram or schema that enables all real-world relations that are required to be represented to be actually represented by database connections that are made possible by the diagram or schema. In order to achieve this, we must understand what is meant by that a database connection represents a real-world relation. This takes a few more notions to define.

#### B. A Database Connection ‘Refers to’ a Real-World Relation

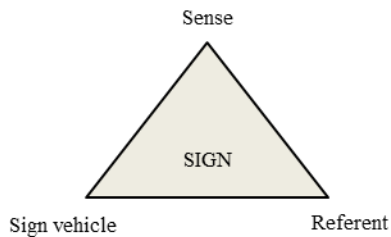


Fig. 1 Peirce’s semiotic triad model [8]

As illustrated in Fig. 1, Peirce’s semiotic triad model shows that the *Representament* (i.e., the form which the sign takes, which is also called ‘sign vehicle’ or ‘signifier’) refers to the *Object* (i.e., the ‘signified’) under the *Interpretant*. The Interpretant is not an interpreter but rather the sense made of the sign. Applying Peirce’s semiotic triad model to databases, a database connection (a ‘sign vehicle’), say  $t$ , refers to a real world relation, say  $s$ , if  $t$  is made up of the entity instances (i.e., nodes in ER instance diagrams in this paper) that refer to the real-world objects involved in  $s$ , and the link in  $t$

refers to the link in  $s$  under the sense-making for database conceptual design. For example, in Fig. 2 below, node  $e_1$  and node  $d_1$  form a database connection, and it refers to the real-world fact that employee  $e_1$  belongs to division  $d_1$ . In such a discussion that a database connection refers to a real-world relation,  $t$  is considered in isolation, i.e., we assume that  $t$  can be and is already ‘picked up’ from the rest of database connections. The reason for this assumption will be made clear shortly.

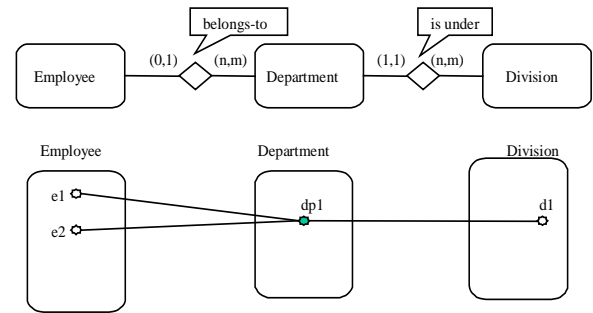


Fig. 2 Database connections shown in an ER diagram

#### C. Relevant and Irrelevant Database Connections

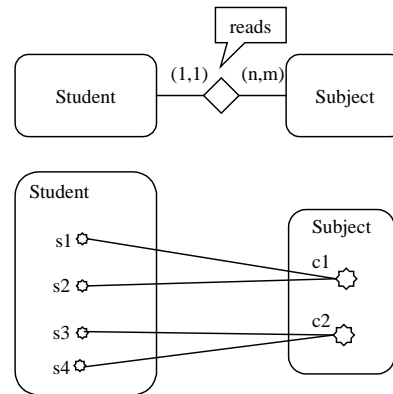


Fig. 3 Irrelevant and relevant schema connections

Due to *nomic structural constraints* [9] that a data model has, an instance of a schema normally has extra connections that come inevitably and ‘for free’. For example, in Fig. 2, path  $(e_2, dp_1, d_1)$  is such a connection, which is resulted in from the existence of path  $(e_1, dp_1, d_1)$  and path  $(e_2, dp_1)$ . These unavoidable and free connections may have nothing to do with what is supposed to be represented. We call such paths *irrelevant database connections* with regard to a particular set of real-world relations. More formally, given a collection of real-world relations  $S$ , a database connection  $t$  is irrelevant to  $S$  if  $t$  refers to no real-world relation in  $S$ , otherwise  $t$  is relevant to  $S$ . Assume that ‘an undergraduate student reads a subject’ is a set of real-world relations. If in Fig. 3, node  $s_1$  refers to a

postgraduate student, then the connection  $(s1, c1)$  is irrelevant to this set of real-world relations.

#### D. Distinguishable Database Connections

A database connection must be distinguishable from the rest in order for it to be useful in terms of representing what it is supposed to represent. Let  $schema1$  be a relational schema or an ER diagram,  $t$  a database connection made possible by  $schema1$ ,  $T$  a type of database connections of which  $t$  is an instance,  $S$  a set of real-world relations of which  $s$  an instance; and let  $t$  refers to  $s$ .  $t$  is distinguishable regarding  $S$  if  $T$  can be explicitly defined by using whatever is available in  $schema1$ . If all irrelevant schema connections can be explicitly defined by whatever is available in  $schema1$ ,  $T$  can also be explicitly defined as a consequence.

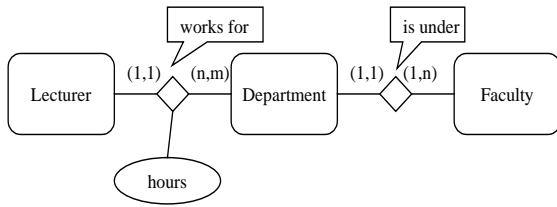


Fig. 4 Relevant database connections that can be explicitly defined

For example, for Fig. 4, assume that only full time lecturers belong to a faculty, and they belong to the faculty under which the department they work for is. With regard to the real-world relation ‘a lecturer belongs to a faculty’, all database connections referring to a part time lecturer and a faculty that are made possible by the path are irrelevant ones. Of all the possible database connections, as long as those that refer to ‘a full time lecturer belongs to a faculty’ can be defined by, say, the post of a lecturer, the hours per week they work, etc, then the relevant database connections are distinguishable. That is, a full time lecturer might be defined as:

Full time lecturer =  $\Sigma_{post = FT}$  Lecturer, or  
 Full time lecturer works for Department =  $\Sigma_{hours > 35}$   
 (Lecturer works for Department)

#### E. A Database Connection ‘Represents’ a Real-World Relation

Only when a database connection refers to a real-world relation and it is distinguishable, can then the database connection be used to indicate that the real-world relation exists. In such a case, we call the former represents the latter. More formally, let  $schema1$  be a relational schema or an ER diagram,  $t$  a database connection made possible by  $schema1$ ,  $S$  a set of real-world relations, and  $s$  an

instance of  $S$ .  $t$  represents  $s$  if  $t$  refers to  $s$  and  $t$  is distinguishable regarding  $S$ .

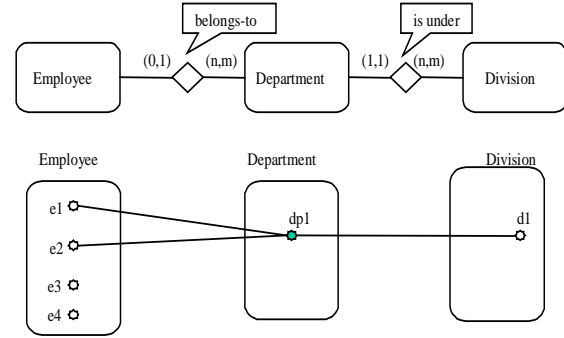


Fig. 5 A database connection is unable to represent a real-world relation due to being indistinguishable

For example, in Fig. 5, which is the same as the one in Fig. 2 where from the discussion earlier database connection  $(e1, dp1, d1)$  is relevant while database connection  $(e2, dp1, d1)$  is irrelevant. Assume that  $e1$  and  $e2$  do not belong to different proper subsets of the entity, then neither  $(e1, dp1, d1)$  nor  $(e2, dp1, d1)$  can be explicitly defined by using, for example, relational algebra or SQL. Consequently the relevant database connection  $(e1, dp1, d1)$  cannot be distinguished from the irrelevant database connection  $(e2, dp1, d1)$ .

It should be noted though that if there is no irrelevant database connection in a path with regard to a type (set) of real-world relations, then the question of whether a database connection is distinguishable does not arise. That is, all database connections represent that set of real-world relations.

The above discussion also shows that a ‘representing’ database connection must be a ‘referring’ one first. But the reverse is not true. Fig. 6 in *higraph* [10] illustrates this point, where  $t$  is a database connection made possible by a database schema,  $S$  is a real-world relation type, and  $s$  is an instance of  $S$ .

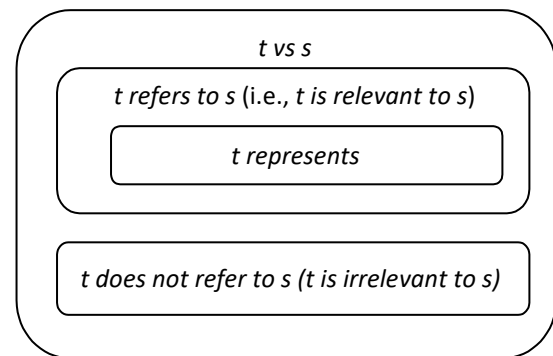


Fig. 6 A database connection refers to or represents a real-world relation

### F. Primary Meaning vs. Implied Meaning of a Path

There are certain types of real-world relation(s) that the database connections of a path can always represent (also refer to, by definition), That is, for such real-world relations, all database connections made possible by the path refer to them, and therefore no irrelevant schema connection is possible. We reveal that such real-world relations are actually the ‘primary meaning’ of a path. In other words, we define ‘primary meaning’ of a data construct [11] in this semiotic way.

For a path in an ER diagram, or two or more relations in a relational schema, a database connection made possible by the path or relational join always has a primary meaning. For example, the path in Fig. 7 has the primary meaning that a lecturer delivers a lecture, and a student attends a lecture. These are the real-world relations that the database connections can always represent.

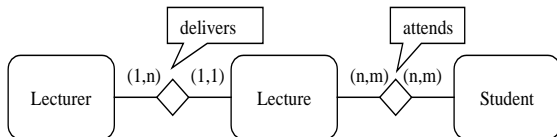


Fig. 7 Primary meaning vs. implied meaning of a path

With certain conditions on both the syntactic level and the semantic level, a database connection may represent a real-world relation that is beyond its primary meaning. For example, the path in Fig. 7 is capable of representing ‘a lecturer lectures a student’, in addition to the primary meaning that we have just said. All such real-world relations constitute the ‘implied meaning’ of a path.

For the conditions on the semantic level, we look at business rules and the logic of a matter. If a lecturer delivers a lecture, and a student attends the lecture, then the lecturer lectures the student. This is logical. In an organization, there might be a business rule, namely ‘an employee may only work on a project that is controlled by the department to which the employee belongs’. Then from ‘an employee works on a project’ and ‘a project is controlled by one (only one) department’, we get ‘an employee belongs to a department.’

For the conditions on the syntactic level, we look out for the structure of a path. Due to its particular structure, a path may not be able to provide database connections that refer to a given set of real-world relations, or a path is capable of providing referring database connections but they are not distinguishable. We pay attention to the length of the path, the participation constraints of the entities, and so on. When the length of a path is greater than one, we watch out for those situations where the ‘plurality of joins’ [12] may apply. Here we examine the concept of ‘plurality of joins’

from the viewpoint that a database connection represents a real-world relation, and extend this concept to cover a more general type of database connections. This would hopefully show as an example how we may approach the representation capability of a database.

### G. The Notion of ‘Plurality of Joins’ Reviewed and Extended

Codd puts forward the concept of ‘plurality of joins’ to explain connection traps in a relational schema [12]. For Codd, given two relations R and S, if there are more than one ternary relation U such that  $\pi_{12}(U) = R$  and  $\pi_{23}(U) = S$ , then R and S have the ‘plurality of joins’. For us, more than one U means that more than one set of database connections meet the above criterion (i.e.,  $\pi_{12}(U) = R$  and  $\pi_{23}(U) = S$ ) and therefore can be established. They are all legitimate syntactically. For example, following Codd [12], we show two joinable relations R and S in Fig. 8, and three different joins of R and S in Fig. 9, Fig. 10, and Fig. 11 respectively below.

R		S	
supplier	part	part	project
1	1	1	1
2	1	1	2
2	2	2	1

Fig.8 Two joinable relations

R*S		
supplier	Part	project
1	1	1
1	1	2
2	1	1
2	1	2
2	2	1

Fig. 9 The natural join of R with S

supplier	Part	project
1	1	2
2	1	1
2	2	1

Fig. 10 Another join of R with S

supplier	Part	project
1	1	1
2	1	2
2	2	1

Fig. 11 Yet another join of R with S

However, not all joins above represent real-world relations except the ‘primary meaning’ (what this means was revealed earlier) of the two entities and the relationship between them. Unless a set of real-world relations happens to be matched by the natural join of R and S, at least one database connection does not refer to any of the set of real-world relations. As we said earlier, such a database connection is called an irrelevant database connection. For example, suppose that only (1,1,2), (2,1,1) and (2,2,1) refer to real-world relations, namely ‘supplier 1 supplies part 1 to project 2’, etc., then (1,1,1) and (2,1,2) are irrelevant database connections. Provided that relevant database connections cannot be explicitly defined (we described this point earlier), a path that is capable of giving rise to ‘plurality of joins’ will not be able to represent a set of real-world relations that involves all the entities in the path and that is not the primary meaning of the path. For the above example, the result of a join cannot be used to represent the real-world relation that ‘a supplier supplies a part to a project’.

This type of situations does not only occur to ‘joinable’ relations [12]. Given two binary relations R and S, as long as  $\pi_{21}(R)$  and S are not functions, that is, they are of many:?:many where ‘?’ stands for one or many, for those tuples  $\sum_{\pi_2(R)=\pi_1(S)}R$  and  $\pi_{23}(U) = \sum_{\pi_1(S)=\pi_2(R)}S$ , the same situation occurs. That is, if we let  $R' = \sum_{\pi_2(R)=\pi_1(S)}R$ , and  $S' = \sum_{\pi_1(S)=\pi_2(R)}S$ , then R’ and S’ will be joinable and therefore have the ‘plurality of joins’. This would result in the database connections of R and S being unable to represent a set of real-world relations that involves all the entities in the path provided that the real-world relations are not the primary meaning of the path<sup>1</sup>. Thus we propose to extend the idea of ‘plurality of joins’ to cover any two relations, say R and S, that can have at least one common element in their common column; and to cover a path of length  $>2$  where at least one instance of the entity in the middle of the path can participate in both relationships in the path. That is, given two relations R and S, if there can be more than one ternary relation U such that  $\pi_{12}(U) = \sum_{\pi_2(R)=\pi_1(S)}R$  and  $\pi_{23}(U) = \sum_{\pi_1(S)=\pi_2(R)}S$ , then R and S have ‘plurality of joins’ (extended from Codd’s definition mentioned earlier). Here a U can also be seen as a set of database connections from the 1st column of R to a common element of the common column of R and S, and then to the 2nd column of S. A similar definition of extended ‘plurality of

joins’ for a path in an ER schema can also be formulated.

#### H. A Database Connection ‘Accurately Represents’ a Real-World Relation

The above definition of representation does not guarantee that a representation is accurate in the sense that what is represented is actually true. For example, a distinguishable little red circle on a map refers to a school thus it represents the school, however the school is now a club and the map is out of date. Such a representation is not accurate. A database connection may also be an inaccurate representation, when, for example, the database is out of date.

Thus, based on the afore-discussion on what is meant by ‘a database connection represents a real-world relation’, now we draw upon Barwise and Seliman’s formulation of ‘representation’ [5, p.235] to define the notion of accurately representation. For the brevity of the presentation, in the rest of the paper, we use ‘path’ in a database model to mean any database connection when a database is viewed conceptually as a graph.

The notion of *accurately representation* can be defined as follows: A path say *PathA* in a conceptual database schema or diagram, e.g., an ER diagram, accurately represents a set of real-world relations say *RelA* if for a given instance of a real-world relation, there is at least one distinguishable instance of a path in the database schema or diagram that refers to the given instance of the real-world relation such that the instance of a path is of *PathA* and the instance of a real-world relation is of *RelA*. In other words, the former represents the latter and the latter is indeed of *RelA*. And furthermore, this applies to all possible instances of *RelA*.

Thus far, we have identified what constitutes the representation capability of a database construct, which is generalized as a path when a database is viewed conceptually as a graph. The sum of such representation capability is that of the database as a whole. Enabled by the representation capability, all the real-world objects that can be represented by constructs of the database constitute the *representation capacity* of the database.

In the sections that follow we wish to explore the representation capability of a database further by looking at informational relationships between database connections and real-world relations. To this end the notion of the ‘information content’ of a sign, an event, and in the most general terms, a state of affairs is relevant.

### III. THE NOTION OF ‘INFORMATION CONTENT’ OF A STATE OF AFFAIRS

Let us consider the following list:

<sup>1</sup> This conclusion is true under the normal condition, namely the natural join of R and S does not happen to refer to the set of real-world relations and the relevant database connections cannot be explicitly defined.

Example 1. That there is smoke carries the information that there is a fire.

Example 2. That he is awarded a grade 'A' for his Programming course contains the information that Jack Brown has gained 80% or above for that course.

Dretske [4, p. 45] defines the nuclear sense of the term 'information content' as follows:

*A state of affairs contains information about X to just that extent to which a suitably placed observer could learn something about X by consulting it.*

Following Dretske, we take information as in the form of 'de re', rather than 'de dicto', that is, in the form of 'a's being F carries the information that b is G'. Dretske [4, p.65] establishes the following definition:

*Information Content: A signal r carries the information that s is F = The conditional probability of s's being F, given r (and k), is 1 (but , given k alone, less than 1).*

In this definition,  $k$  stands for prior knowledge about information source  $s$ . Dretske's approach, which we will extend for our purposes, is based upon the notion of probability [5, pp. 14-18], which is concerned with characterizing events, we first give a definition of event:

**Definition 1.** Let  $s$  be a selection process under a set  $C$  of conditions,  $O$  the set of possible outcomes of  $s$ , which are called states, and  $E$  the power set of  $O$ ,  $X$  is an event if  $E \ni X$  and there is a probability of  $X$ , i.e.,  $P(X)$ .

The notion of 'probability distribution' applies only within a probability space.

**Definition 2.** Let  $s$  be a selection process under a set  $C$  of conditions,  $O$  the set of possible outcomes of  $s$ ,  $E$  the power set of  $O$  and  $E \ni X_i$  for  $i = 1, \dots, n$ ,  $P_s$  is the *probability space* of the events  $X_i$  for  $i = 1, \dots, n$  if  $P_s = \{P(X_1), P(X_2), \dots, P(X_n)\}$  and  $\sum P(X_i) = 1$ .

The information content is concerned with two different levels, namely *tokens* or particulars namely individual things and their *types* [5, p. 69]. It is particulars, i.e., individual things in the world that carry information [5, p. 27]. The information that tokens carry is in the form of types [5, p. 27]. Thus we need a definition for the term particulars of an event.

**Definition 3.** Let  $s$  be a selection process under a set  $C$  of conditions,  $X$  an event concerning  $s$ ,  $X_i$  an instance of  $s$ ,  $X_i$  is a *particular* of  $X$  if  $X_i$  is in a state  $\Omega$ , written  $\Omega = \text{state}(X_i)$ , and  $X \ni \Omega$ .

For example,  $s$  could be concerned with data values going into an attribute, say, the Emp\_Name column of a relational table;  $X_i$  is a data value in the Emp\_Name column at a time  $t$ , which happens to be 'tony\_wu'; the state of  $X_i$ , i.e.,  $\text{state}(X_i) = \text{'a value in Emp\_Name column being tony\_wu'}$ , which is  $\Omega$ ;  $X$  is the disjunction of two states, namely,  $\Omega$  and say,  $\Gamma = \text{'a value in Emp\_Name column being shirley\_wu'}$ . Then,  $X_i$  is a particular of  $X$ .

Given the above concerning the two levels for information content, it would seem appropriate that the above definition of 'the information content of a state of affairs' by Dretske [4, p.65] should be modified as follows.

**Definition 4.** Let  $s$  be some selection process or mechanism the result of which is reduction of possibilities, and therefore be an information source, and  $k$  prior knowledge about  $s^2$ ;

Let  $r$  be an event, and  $r_i$  a particular of  $r$  at time  $t_i$  and location  $l_i$ ;

Let  $s$ 's being  $F$  be an event concerning  $s$ , and  $s_j$  some particular of  $s$ 's being  $F$  at time  $t_j$  and location  $l_j$ ;

$r_i$  carries the information that there must be some  $s_j$  existing at time  $t_j$  and location  $l_j$ , that is, the state of affairs that  $s$  is  $F$  at  $t_j$  and  $l_j$ , if and only if the conditional probability of  $s$ 's being  $F$  given  $r$  is 1 (and less than 1 given  $k$  alone).

**Definition 5.** That a particular  $r_i$  carries the information that a particular  $s_j$  exists can also be termed that the *information content* of  $r_i$  includes  $s_j$ , or in other words,  $s_j$  is in the information content of  $r_i$ .

#### IV. 'INFORMATION CONTENT INCLUSION' RELATION (IIR)

Closely following the previous section, given two events, say  $X$  and  $Y$ , there might be a special type of relations between them, i.e., 'the particulars of event  $Y$  are in the information content of the particulars of event  $X$ '. For brevity, we will also call such a relation 'event  $Y$  is in the information content of event  $X$ '. We suggested calling such relations 'information content inclusion relation' (IIR) [13]. Interestingly it happens that this term also appears in the literature, for example, in her manuscript, Duží [14] points out that information content inclusion relations (in relation to attributes) are of partial order.

**Definition 6.** Let  $X$  and  $Y$  be an event respectively, there exists an *information content inclusion*

---

<sup>2</sup> Note that  $k$  here goes only as far as what counts as a possibility involved in  $s$ , and it is not concerned with whether an observer is able to learn and actually learns something about  $s$  by consulting something else such as  $r$ .

*relation*, IIR for short, from  $X$  to  $Y$ , if every possible particular of  $Y$  is in the information content of at least one particular of  $X$ .

An event may have information content inclusion relation (IIR) with more than one other event. Every one of the latter provides the former with its set of particulars, the whole collection of which is ‘what a suitably placed observer could learn by consulting’ the particulars of the former by following Dretske’s definition [4, p.45] cited earlier. Therefore, this is the information content of the former. That is to say, the information content of an event is the set of events with which the former has an information content inclusion relation.

**Definition 7.** Let  $X$  be an event, the information content of  $X$ , denoted  $I(X)$ , is the set of events with each of which  $X$  has an information content inclusion relation.

Therefore,  $I(X) \ni Y$  is an expression that denotes that event  $Y$  is in the information content of event  $X$  through the particulars of event  $Y$  being in the information content of the particulars of event  $X$  (For the notion of ‘information content’, see Definitions 4 and 5 above). For the sake of the completeness of the definition, we allow  $I(X) \ni X$ , which is a trivial case of  $I(X) \ni Y$ , when  $X$  and  $Y$  are not distinct. Note that in this paper we concern ourselves with the ‘information content inclusion’ relation as just defined only between events (and their particulars), not any other things. This is because we observe that this event-based approach to looking at databases is helpful.

## V. FURTHER FORMULATING REPRESENTATION CAPABILITY OF DATABASES WITH IIR

Now we explore how the representation capability of a database may be further formulated by means of IIR in order to obtain further insight about this concept.

### Proposition 1

Suppose that there is a path  $PathA$  in a database model/schema and there is a real-world relation  $RelA$ , the existence of IIR:  $I(PathA) \ni RelA$  is a sufficient and necessary condition for  $PathA$  to accurately represent  $RelA$ .

### Proof

We prove the ‘sufficient’ part of the above condition by contradiction. Given  $I(PathA) \ni RelA$  as a premise, then by the definition given above every  $r \in RelA$  is in the information content of at least one  $p \in PathA$ , which means whenever a distinguishable instance  $p$  of a path happens to be of  $PathA$ , an instance  $r$  of a real-world relation is of

$RelA$ , otherwise  $r$  may not be of  $RelA$ . And this applies to every  $r \in RelA$ . Now let us assume that  $PathA$  does not accurately represent  $RelA$ . Then it must be the case that there is at least one  $r \in RelA$  such that either no instance  $p$  of a path such that  $p$  represents  $r$  (i.e., either it does not refer to  $r$  or it does but it is not distinguishable) or  $p$  represents  $r$  as being of  $RelA$ , but in fact  $r$  is not of  $RelA$ . This contradicts the premise.

We now prove the ‘necessary’ part of the above condition, also by contradiction. Given that  $PathA$  accurately represents  $RelA$  as a premise, then by the definition given above, for a given instance of a real-world relation, there is at least one distinguishable instance of a path in the database model/schema that refers to the given instance of real-world relation (i.e., the former represents the latter) such that the instance of a path is of  $PathA$  and the instance of a real-world relation is of  $RelA$ . This applies to all possible instances of  $RelA$ . Now let us assume  $I(PathA) \not\ni RelA$ . Then it must be the case that there is at least one instance of  $RelA$  such that it is not in the information content of any instance of  $PathA$ . This means that there must be at least one instance  $r$  of  $RelA$  such that there is no any instance  $p$  of a path such that when  $p$  is of  $PathA$   $r$  is of  $RelA$ . This contradicts the premise.

## VI. CONCLUSIONS

In this paper a seemingly important and yet elusive concept of the *representation capability* of databases has been investigated. The work presented here draws on semiotics, the semantic theory of information presented by Dretske [4] and the information channel theory by Barwise and Seligman [5]. It was found that to approach this concept, to explore and identify what is meant and required by that a database connection *refers to*, *represents* and *accurately represents* a real-world relation respectively is effective. It was also found that that the information content of a database connection includes a real-world relation is a sufficient and necessary condition for the former to be able to accurately represent the latter. All these make the concept of representation capability of a database approachable and definable. Furthermore, based on the representation capability of a database, the *representation capacity* of the database can be defined as well, which is all the real-world relations that can be represented by the constructs that are made possible and available by the database.

## REFERENCES

- [1] S. Gregor, “The Nature of Theory in Information Systems,” *MIS Quarterly*, Vol. 30, No. 3 (Sep., 2006), pp. 611-642 2006



- [2] R. Watson, "Research in Information Systems: What We Haven't Learned," *MIS Quarterly*, (25:4), 2001, pp. v-xv
- [3] R. Weber, "Editor's Comments," *MIS Quarterly*, (27:2), 2003, pp. iii-xii
- [4] F.I. Dretske, *Knowledge and the flow of information*. Blackwell, Oxford, 1981
- [5] J. Barwise and J. Seligman *Information flow: the logic of distributed systems*. Cambridge University Press, London. ISBN 0-521-58386-1, 1997
- [6] R. Stamper, "Organisational semiotics," in: *Information Systems: An Emerging Discipline?*, J. Mingers and F. Stowell, ed., McGraw Hill, London, 1997, pp. 267-283.
- [7] P. B. Andersen, *A Theory of Computer Semiotics*, Cambridge University Press, Cambridge, 1997.
- [8] K. Siau and Y. Tian, "A semiotic analysis of unified modeling language graphical notations," *Requirements Eng* (2009) 14:15–26, Springer-Verlag London Limited, DOI 10.1007/s00766-008-0071-72009
- [9] A. Shimojima, *On the efficacy of representation*. PhD thesis, Indiana University, Bloomington, IN, USA, 1996
- [10] D. Harel, "On Visual Formalisms," *Com. of ACM*, Vol.31, No.5, 1988, pp. 514-30.
- [11] J. Mingers, "Information and meaning: foundations for an inter-subjective account," *Info Systems J*, Vol. 5, 1995, pp. 285-306.
- [12] E.F. Codd, "A Relational Model of Data for Large Shared Data Banks," *Com. of ACM*, Vol. 13 No.6, 1970, pp.377-387.
- [13] J. Feng, "An Analysis of Some Connection Problems in an EER Schema by using a Set of 'Info Concepts'," *Computing and Information Systems*, University of Paisley, UK, M. Crowe ed., Vol.5 No.2, 1988, pp. 63-72.
- [14] M. Duží, "Logical foundations of conceptual modelling," In: *VŠB-TU Ostrava*. 2001

### Authors' Profiles

**Junkang Feng** BSc, MPhil, PhD, FHEA was born in Shanghai China and studied at the Shanghai High School in Shanghai and then graduated from the Institute of Military Engineering of the People's Liberation Army (PLA), the Chinese armed forces, majoring in Guided Missiles Engineering. In China, he lectured at the National University of Defence Technology of the PLA and then worked as a Research Engineer in a research institute of the Shanghai Academy of Spaceflight Technology. In the UK, he received his MPhil from the University of Portsmouth and PhD from the University of the West of Scotland (the UWS) both in Computer Science. He worked as a Research Associate at the University of Manchester UK, a Part-time Lecturer at the University of Portsmouth UK, and a Lecturer at the University of the West of Scotland. Currently Dr Feng is a Senior Lecturer and leads the Database Technology Research Group of UWS. He is also a visiting professor of Donghua University and Beijing Union University in China. Dr Feng's research interests include qualitative information and information flow theories, distributed information systems and database theory and technology, and has published widely in journals, conferences and books in these fields, altogether over 229 items thus far. Dr Feng has completed or been involved in 9 externally (some from EU and China) funded research projects. His work was submitted to RAE (Research Assessment Exercise) 2001 in the UK and REF (Research Excellence Framework) 2014 in the UK. He was Convener of Central Scotland of UK Academy of Information Systems. He reviewed funding applications for the British Council, research manuscripts for the *Journal of Knowledge and Information Systems*, *Journal of Systems, Information Systems Frontier*, *Systemist* and various *WSEAS Transactions*, and conferences of *UKAIS* and *ICISO*. He is a Senior Associate Editor of the *International Journal of Systems and Society*. He has supervised so far 10 PhD programmes (2 as 2nd supervisor) to successful completion and 43 MSc dissertation projects to successful completion