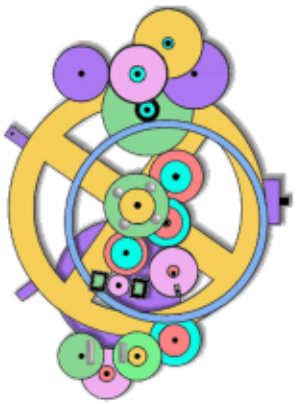
Business Database Design

Classnotes from the Lyceum

The Acropolis - Athens, Greece

Frank Oberle



BUSINESS DATABASE DESIGN

Class Notes from the Lyceum

This book is a follow-up to *Business Database Triage* (see the bibliography beginning on page 113), and provides detailed guidelines for logically approaching the design of relational databases built to support the day-to-day operations of many businesses.

While *Business Database Triage* concentrated primarily on describing and providing examples of the plethora of ill effects that can result from illogical database designs, it only briefly introduced the fundamentals of logical design and a few techniques for implementing the results with modern technology.

This book's subtitle, "Class Notes from the Lyceum," reflects the importance of classic data organization principles first outlined by Aristotle more than two millenia ago for supporting the advance of science, but which are equally invaluable in achieving logical and extensible database designs using modern technology.

Database Design is the process of collecting and organizing the facts of interest to our businesses according to long-established logical and scientific principles, and is quite different from the process of modeling and building databases. This is not, therefore, yet another book about database modeling and construction.

Frank Oberle

Informed by the teachings of Aristotle, Lewis Carroll, Bertrand Russell, and other renowned data organization experts, none of whom ever used a computer, much less a relational database.

Business Database Design – Class Notes from the Lyceum

Publication Information:

Antikythera Publications
www.AntikytheraPubs.com
e-Mail: antikythera@rcn.com

ISBN-10: 069232948X
ISBN-13: 978-0692329481

Preface

An Academic sees a friend … and says, "I heard you died!" The friend says, "Well, you can see I'm alive." And the academic says "Yeah, but I trust the guy who told me more than I trust you." [2]

Continue reading, and don't be that Academic!

"You're neither right nor wrong because people agree with you. You're right because your facts and your reasoning are right."
– Warren Buffet [3]

It seems safe that anyone reading this is related, at least peripherally, to the oddly named[4] Information Technology community (hereinafter referred to by one of its common acronyms "I/T"), and is therefore acutely aware that, quite regularly, some new survey or poll reminds us that: a) I/T projects, particularly applications, take longer to develop and cost more than expected, b) don't accomplish what they were intended to, and c) these same applications become increasingly difficult to maintain and enhance over time.

Despite an almost continuous stream of purported improvements in the tools we use and the often dramatic increases in the capabilities of our processors and other hardware, the assessment of this situation doesn't seem to have changed much from the first such surveys taken in the mid-1980s to the present. We should wonder about that!

Developers – who for the most part would rather be breaking new ground – typically spend far more time than they would like in code maintenance and finding ways of force-fitting new and often unanticipated requests into their existing systems.

2 Quoted in "The 2,000-Year-Old Joke;" Steve Minsky's Anti-Gravity column; Scientific American magazine; October 2014 issue. Determining who to trust can be quite difficult.

3 Quoted in "Tap Dancing to Work: Warren Buffett on Practically Everything 1966-2012;" Fortune Magazine book edited by Carol Loomis

4 I'll comment on why that naming is "odd" later on, but have a suspicion that most practitioners can hazard a guess.

The theory that Date and others wish to teach us, then, must reasonably consist of enhancing the relational model through applying more of the proven underpinnings of logical data organization (i.e. those that Codd didn't consider immediately relevant) to the relational model, or enhancing the relational model to take advantage of the technical progress since Codd first presented his proposals over forty-five years ago.

As children, we are often taught, rightly or not, that "Columbus 'discovered' America," raising a number of questions: hadn't the people he encountered on landing previously 'discovered' America? Didn't many earlier explorers 'discover' America? Did any of them realize that what they 'discovered' was 'America?'

Over the intervening years, many have claimed refinements and even improvements to the Relational Model, but we need to be quite wary of these. Certainly, there is significant room for improvements in how the science of data organization is applied to technology; the technology itself has obviously improved since Dr. Codd unveiled his discovery. We need to be particularly wary when encountering statements that actually conflict with the aforementioned underpinnings. Date, for instance, now offers us "The Third Manifesto" … So, who put Date in charge of the Relational Model?

Those claiming to improve the underlying science, however, have a significantly higher standard to meet, and most such claims don't come close – either distorting the proven tenets underlying the model, or simply ignoring them for some convenience or another.

A 1993 interview with Dr. Codd:

DBMS Magazine: "Where did [the term] 'normalization' come from?"

Codd: " … then President Nixon was talking a lot about normalizing relations with China. I figured that if he could normalize relations, so could I."

Yes, Virginia, Codd had a sense of humor!

Of course, when discussing Relational Databases (however anyone attempts to define them), we encounter the mysterious subject of Normalization, associated in the minds of most developers with the "Normal Forms."

Codd was, in fact, the first to define what we know as Normal Forms,[20] but he decidedly didn't invent (or even need to) the concept and definition of Normalization, as the often quoted 1993 interview[21] he gave is sometimes incorrectly interpreted. In the section "Normalization: Carroll, Codd, and Nixon" on pages 93 and 94 of *Business Database Triage* suggests:

20 At least as far as I have been able to determine.

21 Codd [3]; see page 115 for information.

The average Business is not likely to store information in their database about individual triangles or pencils – two of the subjects used as examples – making these seemingly odd choices for our modeling examples. Given that choosing common entities such as Employee or Customer would provide a number of distractions due to our ingrained prejudices, habits, and preconceptions about how these entities have been traditionally (mis)handled, the use of such unrealistic entities will actually make it a more useful analysis exercise.

Suspension of Disbelief takes two primary forms in this discussion – what I have titled "Ceci n'est pas une Pipe" and "Anthropomorphism," terms that will be explained as they are discussed below.

Suspension of Disbelief I – Ceci n'est pas une Pipe

Consider the common, and easily recognized object shown on the right. If asked "What is this Thing?," what would your response be?

The most common answer, of course, would be "it's a Pencil," which is certainly a colloquially acceptable response in normal conversation.

A less common and more general response might be "it's a writing implement," or something similar. As we know from Aristotle[40], that is an even less satisfactory answer because it is not the most precise we could give.

Of course, in any group of more than two or three persons, the "wise a**" response would be something along the lines of "it's a picture of a pencil." In point of fact, this is the most correct, precise and accurate answer.[41] It is, in fact, a picture.

In the opening sentence of his Categories, Aristotle's tells us that "a real man and a figure in a picture can both lay claim to the name 'animal'." The Belgian surrealist René Magritte gives us an illustration of an important lesson in analyzing "Things." Let's assume that we had defined the noun "pipe" sufficiently to avoid confusion with plumbing artifacts; we recognize, after all, that "pipe" is a homonym, but there isn't really a good alternative word we can choose. If we were to attempt a formal categorization of the various types of "Pipe," we might get lucky and have a representative selection of

40 [REFERENCE]

41 Wise a** but intelligent nitpickers, when properly managed, are the general class of people … [continue with data organizers and database designers]

actual physical examples to examine. We would then proceed with all the steps we will soon learn in order to accomplish the analysis needed to create a logical data structure.

But it is far more likely in many business situations that we won't have that luxury; it is even likely in some cases, that such access might even prove to be confusing (if, for instance, you are not a Subject Matter Expert – an SME – which in this case probably means a pipe smoker). The truth is that you would likely be able to glean as much if not more information from a good pipe-smokers' catalog.

We need to be careful to avoid too slavish precision, and we also need to be aware that "denials" come in various forms – some actionable and some not. In this example Magritte gives us what Aristotle refers to as a Contrary: "This is not a Pipe." While this puts us on notice that we need to pay close attention during analysis it doesn't really give us anything to evaluate. As we know from Aristotle, a Contrary is a less satisfactory denial because it is not the most precise we could give.

More to the point when we're discussing Logic, a negative assertion cannot be proven.

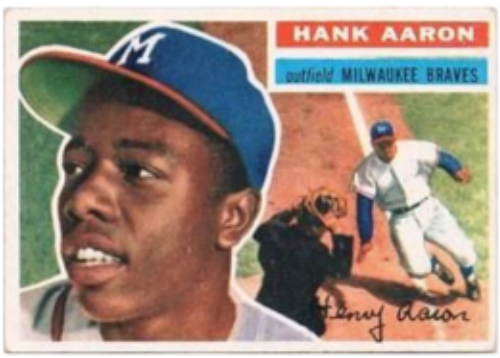
A stronger and therefore more actionable form of denial is what Aristotle refers to as a Contradictory, which gives us a different assertion and, more importantly, one that we can evaluate.

All other things being equal, the "wise a**"es in our IT departments are generally better suited to the sort of data analysis required for the logical design of relational database structures than their more boring colleagues. Nonetheless, we need to suspend our disbelief and treat many "illustrations," as well as many "placeholders," as if they were the actual object depicted. We must also consider whether there are situations where this suspension of disbelief is not appropriate. Pay Attention.

In his 1928 painting titled "The Treachery of Images,"[42] the artist René Magritte painted the caption "Ceci n'est pas une pipe" (In English, "This is not a pipe") on the image itself, and commented that all one needed to do to prove this was to attempt filling its bowl with tobacco and lighting it. For our purposes, however, this literal interpretation won't do. Although the "wise a**"es in our IT departments are, all other things being equal, generally better suited to the sort of data analysis required for the logical design of relational database structures, we will need to suspend

The Treachery of Images
René Magritte - 1928

42 Currently in the collection of the Los Angeles County Museum of Art.



Illustrating how easy it can be to become "boxed in" and non-extensible.
Unraveling Perspectives and making the best Choice
The Role of Domain Expertise
The Role of Subject Matter Experts (SMEs)

4

4 - Design Exercise: The Trading Card Collection

Introduction

You've just joined a new corporation that prides itself on having a rigorously designed operational database which, your new immediate supervisor claims, is fully normalized, very well constrained, and with no duplicate data elements. It is, says the supervisor, a relational database built with Microsoft SQL-Server™.

As it happens, the CEO's young son, an avid sports enthusiast – although not himself athletic – has begun a new hobby. The CEO has asked for a crack database designer to build a relational database to support managing the boy's collection. You, being the newest member, have been "volunteered" to keep the CEO happy. And your supervisor doesn't hide the fact that, in addition to keeping the CEO out of the department's hair, this will provide him with an opportunity to evaluate your understanding of relational technology and design techniques before turning you loose on the company's "real" database. You are told to treat this as an enterprise-class database implementation. No four column spreadsheet here!

You meet with the CEO and his son, who pulls out what appears to be a pile of eight pieces of thin cardboard, each about 2.5 inches by 3 inches, that he assures you are valuable investments that have steadily increased in value over more than fifty-five years. The child explains that this is the beginning of what he intends to be a world class collection that will one day pay for his college education.

He turns the cards over and then you see their colorful faces – pictures of baseball players. It's easy to see that the boy has begun a collection of baseball cards. You might even notice that all the players are New York Yankees.

The cards are illustrated on the right. From left-to-right, and top-to-bottom, they are:

- a 1957 Mickey Mantle card
- a 1957 Elston Howard card
- a 1956 Mickey Mantle card
- a 1956 Whitey Ford card
- a torn 1956 Yogi Berra card
- a good 1956 Yogi Berra card
- a 1957 Don Larsen card
- a 1957 Yogi Berra card

So, what's the proper approach to organizing the data needed to track this collection?

We can view this physical grouping of cards in several ways. It might, for instance, be considered a single Collection of eight Baseball Cards, which happens to be a Set of seven New York Yankee Baseball Cards with one duplicate card. It might be considered two Collections of four Baseball Cards each: … a set of four 1956 Yankee Baseball Cards with one duplicate and a set of four 1957 Yankee Baseball Cards (with no duplicate cards.)

Because, so far as we know at this point, there will need to be a central table containing each piece of the collection, we know we need to accomplish a few things to start – first and foremost answering the question "what are these Things?" – in both very specific and very general senses.

What is the most general description that applies to every piece of the boy's collection as well as the collection as a whole? _____*(fill in)*_____.

What is the most specific description that applies to every piece of the boy's collection as well as the collection as a whole? _____*(fill in)*_____.

Using a modeling tool, we create an Entity-Set and begin building a list of attributes. Based on the previous two questions, we begin by choosing an appropriate name for what will become the central Table: ____*(fill in)*_____.

The Name, remember, should make perfect sense when forming Propositions such as:

- The torn 1956 Yogi Berra Card is a [Your-Chosen-Table-Name].
- The 1957 Don Larsen card is a [Your-Chosen-Table-Name].

… and that this and similar Propositions for each and every other card in the collection are categorically True.

Whether [Your-Chosen-Table-Name] in the Propositions above is "Trading Card," "Sports_Card," "Baseball Card," "Yankee Baseball Card," or something else, you should spend some time and effort identifying the Pros and Cons of each.

You still of course consider this a "working name," knowing that there are a number of steps that need to be taken before committing to a definition of this entity-set much less an appropriate name for it. Although it is generally not a good practice to assume that each entity-set will become a table, this particular structure is central enough to the subject matter that, for this exercise, we won't distinguish between an Entity-Set and a Table.

Now you identify the peculiar adjunct of entity definition (your table) – the set of attributes that make this table what we suspect it is – and indicate those attributes that are therefore Essential. You begin by listing all attributes – potential columns and relationships – you can determine, both from observation and from conversations with the collector himself. As we'll see, we should be rightly suspicious that the collector is actually a Subject Matter Expert.[46] The list, quite informal at this stage, should look something like the following:

Attribute Name	Essential or Accidental ?	Description and/or Constraint(s) and Notes
(fill in)	*(fill in)*	*(fill in)*

46 In this case, of course, the fact that he is a child constitutes an excuse of sorts; with some experience you will learn that more business "experts" than you might expect don't meet any reasonable criteria for that appellation any more than the CEO's young son does. You must develop a sense of when someone is generalizing what should be specific or doing the opposite. Good design requires an understanding of each perspective.

Preliminary Analysis Results

The list you created in the previous section is likely reasonably similar to the Table below, which illustrates many, but certainly not all, possible Attributes. Given the lack of realistic context, exact matches are unlikely, but there should be some similarities.

Attribute Name	Essential or Accidental ?	Description and/or Constraint(s) and Notes
Identifier	Essential	An arbitrary but unique identifier for each Card
Player Name	Essential	
Team Name	Essential	
Year	Essential	
Publisher	Accidental	
Acquisition Date	Accidental	
Acquisition Price	Accidental	
Currency	Accidental (implied)	
Condition	Accidental	
League	Essential	
Sport	Essential	Is this an attribute of League?
	Accidental	
	Accidental	
et cetera …		

Since this table will be the central point of focus for the database we are building, we should ensure that it meets all the criteria Aristotle prescribed for a Class.

A little preliminary analysis tells us several things about this collection:

Each element is uniquely identified, but not necessarily unique. - the missing qualification is no duplicates and unique id ...

A Class must be defined by a Peculiar Adjunct that permits us to determine that something is or isn't a Member of the Class.

So far, Each element discussed above is physical. We can tell the two 1956 Yogi Berra cards apart by observing them side-by-side on the table. An avid collector might even mark their plastic sleeve with an identifier and perhaps a quality rating of some sort. Once we get away from actual "things," we run into a problem, so for data we must identify duplicates in some fashion to tell them apart.

(identified, but unordered, uncategorized, etc.)

For a data model, we would likely add inter-related attributes like value and condition. This possibly implies providing more precise identification of each element.

Counterfeits and Domain Violations

As the table is coming together, the CEO's son appears with two new Yankees cards he acquired from a wandering trader at a show. These are:

- a 1956 Babe Ruth Card
- a 1957 Derek Jeter Card

The lack of any residual Bubble Gum smell doesn't tell us much, since it would likely have dissipated after half a century, but we recall just seeing the celebrations of Jeter's final game a short time ago. This leads us to suspect the provenance of these cards may be questionable. A little research tells us that Babe Ruth died in 1948 and Derek Jeter wasn't born until 1974, confirming our suspicion that our young collector may have just learned a useful lesson.

Establish a domain of valid cards to serve as a constraint. Describe in general terms what the domain might specify: _____

Is it the team, or the printed cards? If the latter, do we need to know who printed the cards? (e.g. Fleer, Topps, and Score were common).

Scope Creep

Now, showing many of the hallmarks of a typical business executive, the wünderkind to whom your career now seems devoted announces that he will be expanding the scope of his collection, and brings in two new cards:

From left to right, the additional New York baseball cards are:

- a 1956 Hoyt Wilhelm Card
- a 1956 Willie Mays Card

But wait, these cards are for New York Giants players, not New York Yankees players like the previous cards. But certainly the addition of Willie Mays – one of the best known players of that era – and the knuckle-ball thrower[47] Hoyt Wilhelm will certainly add to the collection's future value.

47 … and a member of the exclusive fraternity of players who pitched a no-hitter, which he did during a later stint with the Baltimore Orioles.

Subject

The next addition to the boy's collection are two more Willie Mays cards. From left to right, the two new Giants cards are:

- a 1958 Willie Mays Card
- a 1959 Willie Mays Card

Giants last season in New York was 1957

San Francisco, not New York.

Are the players / cards grouped by Team, City, State, Location, Franchise?

Note: Washington Senators moved to Minneapolis and became the Minnesota Twins. (now known by a State name, not a City name, as are the Texas Rangers).

Are teams independent by season, or is there continuity required? League composition? Does this depend on continuity of ownership? Majority of players?

Football Giants left New York, although retained the name. The team's first year in New Jersey was 1975.

Subject

The next pair of cards our intrepid collector showed up with were two more examples of New York Giants cards from the year 1956. From left to right, the additional cards are:

- a 1956 Charley Conerly Card
- a 1956 Frank Gifford Card

Uh Oh ...

Was Sport listed as an attribute? If not, recovery is not that difficult since, although the design didn't completely reflect reality, it didn't contradict reality.

Establish a "sport" domain table; add sport column as nullable; update all sport columns to "baseball"; set sport column to not null; add new football cards.

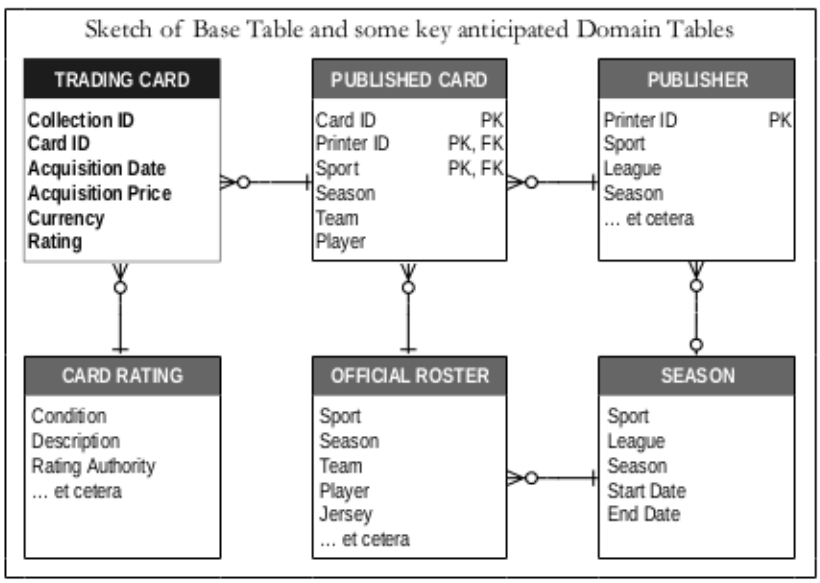
Season versus Year?

How was the attribute for "1956" established? As a Year, or as a Season? What defines a season (a baseball season takes place in one calendar year, although that is actually arbitrary. Several other sports, including football, have seasons that begin in one year and end in another. The NFL championship for the 2013 season was determined by the Super Bowl played on February 2, 2014. The "85 Bears" played that season's Super Bowl in 1986.

Cy Young award winner Doug Drabeck shown with Chicago White Soz in 1998, but was on the Baltimore Orioles that year.

1995 Carlos Beltran's card has Juan LeBron's photo on the card.

Stats per player per season ??



What if a card has multiple ratings from different catalogs? This would require a _____ table between Trading_Card and Card_Rating tables.

A Trigger would immediately exit if it doesn't find at least 2 angle values and 1 side value OR at least 2 side values and 1 angle value.

Ticket stub for Don Larsen's perfect game in game 5 of 1956 World Series.($400)

Does the collection of cards constitute a {Set}, a [Multi-Set], or a Class?

Each element is uniquely identified, but not necessarily unique. As a Set, what we have is seven instances with a duplicate of one instance. It is therefore a Multi-Set.

Assuming this collection is a Set, is it bounded? If so, how?

Number of Array Elements is not dependent on whether any values are duplicates.

Number of Set Elements is dependent, since there is an id and a quantity.

The Set {1,2,2,3,5,7,7,9}

There are eight numbers with two duplicates.

has 8 elements, but 6 members.

And so on ...

Next, the now annoying brat, now having developed all the traits of what the IT community disparagingly refers to as a "User," shows up with several rare vintage Pokemon cards ...

"*Classnotes from the Lyceum*" is written by the author as a followup to "*Business Database Triage*," and provides detailed guidelines for the design of Relational Databases for use in Business.
While Business Database Triage concentrates on determining what effects result from poor and illogical database designs, Classnotes addresses techniques for avoiding these situations by following the scientific and logical approaches to data organization laid out by Aristotle more than two millenia ago.

Antikythera Publications

Databases
Information Management
Business