> *organizations, the effort is so substantial that it is divided into separate roles, where a data modeler is responsible for oneness and a data architect is responsible for sameness.*

Oneness, Sameness, and Categories are tightly intertwined with one another.

ONENESS (WHAT IS "ONE THING"?)

Consider "book." If an author has written two books, a bibliographic database will have two representatives. (You may temporarily think of a representative as being a record.) If a lending library has five circulating copies of each, it will have ten representatives in its files. After we recognize the ambiguity, we try to carefully adopt a convention using the words "book" and "copy." But it is not natural usage. Would you understand the question "How many copies are there in the library?" when I really want to know how many physical books the library has altogether?

There are other connotations of the word "book" that could interfere with the smooth integration of databases. A "book" may denote something with hard covers, as distinguished from things in soft covers like manuals, periodicals, etc. Thus a manual may be classified as a "book" in one library, but not in another. I don't always know whether conference proceedings constitute a "book."

A "book" may denote something bound together as one physical unit. Thus, a single long novel may be printed in two physical parts. When we recognize the ambiguity, we sometimes try to avoid it by agreeing to use the term "volume" in a certain way, but we are not always consistent. Sometimes several "volumes" are bound into one physical "book." We now have as plausible perceptions: the *one* book written by an author, the *two* books in the library's title files (Vol. I and Vol. II), and the *ten* books on the shelf of the library which has five copies of everything.

Incidentally, the converse sometimes also happens, as when several novels are published as one physical book (e.g., collected works).

So, once again, if we are going to have a database about books, before we can know what one representative stands for, we had better have a consensus among all users as to what "one book" is.

Going back now to parts and warehouses, the notion of "warehouse" opens up another kind of ambiguity. There is no natural, intrinsic notion of what constitutes "one warehouse." It may be a single building, or a group of buildings separated by any arbitrary distance. Several warehouses (e.g., belonging to different companies) may occupy the

same building, perhaps on different floors. So, what is "one warehouse"? Anything that a certain group of people agrees to call a warehouse. Given two buildings, they might agree to treat them as one, two, or any number of warehouses—with all perceptions being equally "correct."

IBM assigns "building numbers" to its buildings for the routing of internal mail, recording employee locations, and other purposes. One two-story building in Palo Alto, California, is "Building 046," with the two stories distinguished by suffixes: 046-1 and 046-2. Right next door is another two-story building. The upper story is itself called "Building 034," and the lower story is split into two parts called "Building 032" and "Building 047." IBM didn't invent the situation. The designations correspond to three different postal addresses: 1508, 1510, and 1512 Page Mill Road are all in the same building.

Another IBM location in the hills of San Jose, California, is apparently one building, since it has one building number. The structure has eight distinct towers. Signs inside direct you to "Building A," "Building B," etc. How many buildings are there?

"Street" is another ambiguous term. What is one street? Sometimes the name changes; that is, different segments along the same straight path have different names. Based on a comparison of addresses, we would probably surmise that people on those various segments lived on different streets. On the other hand, different streets in the same town may have the same name. Now what does an address comparison imply?

Is a street terminated by city, county, state, or national boundaries? Suppose the street just ran right across the boundary, same name and all. Would you be inclined to say that people living in different countries lived on the same street?

Does the term "street" imply that motor vehicles can drive on it? Some are narrower than alleys, and some are pedestrian malls.

Does the term "street" include freeways, highways, thruways, expressways, tollways, parkways, autobahns, autopistes, autostradas, autoroutes, dual carriageways, motorways? (I'm really just trying to convey one idea—what do they call it in your neighborhood?) Very often, one highway will coincide with portions of many different streets along its route. Does a highway name count as a street name? Along some segments, the highway name might be the only street name. Various street segments will have various multitudes of names ("look at all the highway markers on that pole!"). And, after I make a turn, whether or not I'm on the "same street" may depend on my own state of

mind: which street name did I think I was following? Finally: if I drive from Illinois to California on Highway 66, have I been on the same street all the way?

Thus, the boundaries and extent of "one thing" can be very arbitrarily established. This is even more so when we perform "classification" in an area that has no natural sharp boundaries at all. The set of things that human beings know how to do is infinitely varied, and changes from one human being to another in the most subtle and devious ways. Nonetheless, the "skills" portion of a personnel database asserts a finite number of arbitrary skill categories, with each skill being treated as one discrete thing, i.e., it has one representative. The number and nature of these skills is very arbitrary (i.e., they do not correspond to natural, intrinsic boundaries in the real world), and they are likely to be different in different databases. Thus, a "thing" here is a very arbitrary segment partitioned out of a continuum. This applies also to the set of subjects in a library file or information retrieval system, to the set of diseases in a medical database, to colors, etc.

This classification problem underlies the general ambiguity of words. The set of concepts we try to communicate about is infinite (and non-denumerable in the most mind-boggling sense), whereas we communicate using an essentially finite set of words. (For this discussion, it suffices just to think about nouns.) Thus, a word does not correspond to a single concept, but to a cluster of more or less related concepts. Very often, the use of a word to denote two different ideas in this cluster can get us into trouble.

A case in point is the word "well," as used in the data files of an oil company. In their geological database, a "well" is a single hole drilled in the surface of the earth, whether or not it produces oil. In the production database, a "well" is one or more holes covered by one piece of equipment, which has tapped into a pool of oil. The oil company had trouble integrating these databases to support a new application: the correlation of well productivity with geological characteristics.

SAMENESS (HOW MANY THINGS IS IT?)

A single physical unit often functions in several roles, each of which is to be represented as a separate thing in the information system. Consider a database maintaining scoring statistics for a soccer team, both on a position basis and on an individual basis. The database might have representatives for 36 things: 11 positions and 25 players. When Joe Smith, playing halfback, scores a goal, the data about two things is modified: the number of goals by Joe Smith, and the number of goals by a halfback. That human figure standing on the field is represented as