DATA AND REALITY

Basic Assumptions in Data Processing Reconsidered

William KENT

IBM

San Jose, California

NORTH-HOLLAND PUBLISHING COMPANY – AMSTERDAM • NEW YORK • OXFORD

1.0 ENTITIES

> "Entities are a state of mind. No two people agree on what the real world view is." [Metaxides]

An information system (e.g., data base) is a model of a small, finite subset of the real world. (More or less -- we'll come back to that later.) We expect certain correspondences between constructs inside the information system and in the real world. We expect to have one record in the employee file for each person employed by the company. If an employee works in a certain department, we expect to find that department's number in that employee's record.

So, one of the first concepts we have is a correspondence between things inside the information system and things in the real world. Ideally, this would be a one-to-one correspondence, i.e., we could identify a single construct in the information system which represented a single thing in the real world.

Even these simple expectations run into trouble. In the first place, it's not so easy to pin down what construct in the information system will do the representing. It might be a record (whatever that means), or a part of one, or several of them, or a catalog entry, or a subject in a data dictionary, or .... For now let's just call that thing a representative, and come back to that topic later. Let's explore instead how well we really understand what it is that we want represented.

As a schoolteacher might say, before we start writing data descriptions let's pause a minute and get our thoughts in order. Before we go charging off to design or use a data structure, let's think about the information we want to represent. Do we have a very clear idea of what that information is like? Do we have a good grasp of the semantic problems involved?

Becoming an expert in data structures is like becoming an

expert in sentence structure and grammar. It's not of much value if the thoughts you want to express are all muddled.

The information in the system is part of a communication process among people. There is a flow of ideas from mind to mind; there are translations along the way, from concept to natural languages to formal languages (constructs in the machine system) and back again. An observer of, or participant in, a certain process recognizes that a certain person has become employed by a certain department. The observer causes that fact to be recorded, perhaps in a data base, where someone else can later interrogate that recorded fact to get certain ideas out of it. The resemblance between the extracted ideas and the ideas in the original observer's mind does not depend only on the accuracy with which the messages are recorded and transmitted. It also depends heavily on the participants' common understanding of the elementary references to "a certain person", "a certain department", and "is employed by".

1.1 ONE THING

What is "one thing"?

That appears at first to be a trivial, irrelevant, irreverent, absurd question. It's not. The question illustrates how deeply ambiguity and misunderstanding are ingrained in the way we think and talk.

Consider those good old workhorse data base examples, parts and warehouses. We normally assume a context in which each part has a part number and occurs in various quantities at various warehouses. Notice that: various quantities of one thing. Is it one or many? Obviously, the assumption here is that "part" means one kind of part, of which there may be many physical instances. (The same ambiguity shows up very often in natural usage, when we refer to two physical things as "the same thing" when we mean "the same kind".) It is a perfectly valid and useful point of view in the context of, e.g., an inventory file: we have one representative (record) for each kind of thing, and speak loosely of all occurrences of the thing as collectively being one thing. (We could also approach this by saying that the representative is not meant to correspond to any physical object, but to the abstracted idea of one kind of object. Nonetheless, we do use the term "part", and not "kind of part".)

Now consider another application, a quality control application, also dealing with parts. In this context, "part" means one physical object; each part is subjected to certain tests, and the test data is maintained in a data base

separately for each part. There is now one representative in the information system for each physical object, many of which may have the same part number.

In order to integrate the data bases for the inventory and quality control applications, the people involved need to recognize that there are two different notions of "thing" associated with the concept of "part", and the two views must be reconciled. They will have to work out a convention wherein the information system can deal with two kinds of representatives: one standing for a kind of part, another standing for one physical object.

I hope you're convinced now that we have to go to some depth to deal with the basic semantic problems of data description.

We are dealing with a natural ambiguity of words, which we as human beings resolve in a largely automatic and unconscious way, because we understand the context in which the words are being used. When a data file exists to serve just one application, there is in effect just one context, and users implicitly understand that context; they automatically resolve ambiguities by interpreting words as appropriate for that context. But when files get integrated into a data base serving multiple applications, that ambiguity-resolving mechanism is lost. The assumptions appropriate to the context of one application may not fit the contexts of other applications.

There are a few basic concepts we have to deal with here:

* Oneness.
* Sameness. When do we say two things are the same, or the same thing? How does change affect identity?
* What is it? In what categories do we perceive the thing to be? What categories do we acknowledge? How well defined are they?

These concepts and questions are tightly intertwined with one another.

Consider "book". If an author has written two books, a bibliographic data base 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 data bases. 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 data base 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 Santa Teresa, 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?

Sometimes a street is made up of discontinuous segments, perhaps because intervening sections just haven't been built yet. They may not even be on a straight line, because the ultimate street on somebody's master plan curves and wiggles all around. And sometimes I can make a right turn, then after some distance make a left turn and be back on a street with the same name as the first. Is that one street with a jog? When do we start thinking of these as different streets having the same name?

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 New York 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 data base 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 data bases. 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 data base, 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 data base, a "well" is a single hole drilled in the surface of the earth, whether or not it produces oil. In the production data base, 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 data bases to support a new application: the correlation of well productivity with geological characteristics.

1.2 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 data base maintaining scoring statistics for a soccer team, both on a position basis and on an individual basis. The data base 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 (and is) two things: Joe Smith and a halfback.

Consider the question of "sameness". Suppose Joe switches to fullback, and scores another goal. Did the same thing make those two goals? Yes: Joe Smith made both. No: one was made by a halfback, the other by a fullback.

Why is that human figure perceived and treated as two things, rather than one or three or ninety-eight? Not by any natural law, but by the arbitrary decision of some human beings, because the perception was useful to them, and corresponded to the kinds of information they were interested in maintaining in the system.

If the file only had data about player positions, then the same physical object would be treated as being different things at different times. Joe is sometimes a halfback and sometimes a fullback. From the perspective of this file, his activities are being performed by two different entities.

Also consider two related people (e.g., husband and wife) who work for the same company. When considering medical benefits, each of these people has to be considered twice: once as an employee, and once as a dependent of an employee. How many people are involved?

Or suppose a person held two jobs with the company, on two different shifts. Does that signify one or two employees? Shipping clerk John Jones and third-shift computer operator John Jones might be the same person. Does it matter? Sometimes.

The notion is also applicable to warehouses. From the point of view of another application, the thing involved is not a warehouse at all, but a building or property on the assessment rolls.

It is plausible (bizarre, perhaps, but plausible) to view a certain employee and a certain stockholder as two different things, between which there happens to exist the relationship that they are embodied in the same person. There would then exist two representatives in the system, one for the employee and one for the stockholder. It's perfectly all right, so long as users understand the implications of this convention (e.g., deleting one might not delete the other).

Transportation schedules and vehicles offer other examples of ambiguities, in the use of such terms as "flight" and "plane" (even if we ignore the other definitions of "plane" having nothing to do with flying machines). What does "catching the same plane every Friday" really mean? It may or may not be the same physical airplane. But if a mechanic is scheduled to service the same plane every Friday, it had better be the same physical airplane. And another thing: if two passengers board a plane together in San Francisco, with one holding a ticket to New York and the other a ticket to Amsterdam, are they on the same flight?

Classification, e.g., of skills, impacts the notion of "sameness" as much as the notion of "how many". The way we

partition skills determines both how many different things we recognize in this category, and when we will judge two things to be the same. Consider a group of people who know how to do such things as paint signs on doors, paint portraits, paint houses, draw building blueprints, draw wiring diagrams, etc. One classifier might judge that there is just one skill represented by all of these capabilities, namely "artist", and that every person in this group had the same skill. Another classifier might claim there are two skills here, namely painting and drawing. Then the sign painter has the same skill as the portrait painter, but not the blueprint drawer. And so on.

The same game can be played with colors. Two red things are the same color. What if one is crimson and the other scarlet?

The perceptive reader will have noticed that two kinds of "how many" questions have been intermixed in this section. At first we were exploring how many kinds of things something might be perceived to be. But occasionally we were trying to determine whether we were dealing with one or several things of a given kind. If you can't apply that distinction to the preceding discussions, then please don't become a data base administrator. I fear your data base may well become a minefield of semantic traps.

For another example of the latter kind, consider program problem reports (known as APAR's in IBM). Considerable effort is often expended in determining that the symptoms reported in two APAR's are caused by the same programming error; thereafter, the two APAR's are considered to be the "same". (The correctness of this view depends on whether you think the entity involved is the programming error or the problem report.)

And analogously, much of the fuss in many insurance claims and court battles revolves around determining whether several things relate to the "same" illness or injury.

1.3 CHANGE

And then there's change. Even after consensus has been reached on what things are to be represented in the information system, the impact of change must be considered. How much change can something undergo and still be the "same thing"? At what point is it appropriate to introduce a new representative into the system, because change has transformed something into a new and different thing?

The problem is one of identifying or discovering some

essential invariant characteristic of a thing, which gives it its identity. That invariant characteristic is often hard to identify, or may not exist at all.

We seem to have little difficulty with the concept of "one person" despite changes in appearance, personality, capabilities, and, above all, chemical composition. (The proportions and structure -- i.e., the chemical formulas -- may not change much, but the individual atoms and molecules are continually being replaced... again illustrating an ambiguity between "same kind" and "same instance": how rapidly is the chemical composition of your body changing?) When we speak of the same person over a period of time, we certainly are not referring to the same ensemble of atoms and molecules. What then is the "same person"? We can only appeal to some vague intuition about the "continuity" of -- something -- through gradual change. The concept of "same person" is so familiar and obvious that it is absolutely irritating not to be able to define it. Definitions in terms of "soul" and "spirit" may be the only true and humanistic concepts, but, significantly, we don't know how to deal with them in a computer-based information system. It is only when the notion of "person" is pushed to some limit do we realize how imprecise the notion is. This is the basis of some legal issues.

Modern medicine is dissecting our concept of "person" via transplanted and artificial limbs and organs. The Hopi Indians consider mental activity to be in the heart [Whorf]; they might argue that the recipient of a heart transplant becomes the person who the donor was -- the donor has merely acquired a new body. (Is it a heart transplant or a body transplant?) We are more likely to take that position with respect to the brain, rather than the heart. A number of legal issues will have to be resolved when brain transplants begin to be performed (and the issues may get more complex if just portions of the brain are transplanted).

In an information system maintaining data about people, we will have to decide which information gets interchanged between two representatives. Which information is to be associated with the body, and which with the brain? A name? A spouse? Other relatives? How is the medical history rearranged? Who has which job? Skills? Financial obligations?

We also have some issues regarding the beginning and ending of a person. It makes sense in the context of some medical records to treat an unborn fetus as an unborn person; observations during pregnancy become a part of that person's medical history. A recent court case considered the question of whether an unborn fetus was eligible for welfare benefits, which would have made the fetus representable in the welfare data base. After death, a person ceases to

exist for many legal purposes, but the data about him (or his body) continues to be relevant to a cemetery, or a coroner, or a medical researcher.

An analogous situation exists with automobiles. Suppose you and I start trading parts of our cars -- tires, wheels, transmissions, suspensions, etc. At some point we will have exchanged cars, in the sense that the Department of Motor Vehicles must change their records as to who owns which car -- but when? What is the "thing" which used to be my car, and when did you acquire it? The Department of Motor Vehicles (at least in California, I believe) has made an arbitrary decision: the "essence" of a car is the engine block, which is (they assume) indivisible and is uniquely numbered. Owning and registering a car is defined to mean owning and registering the engine block. All the other parts of the car can be removed or replaced without altering the identity of the car.

What would happen if another state had a different convention for establishing the identity of a car? Could their two data bases be integrated?

The same kinds of questions apply to organizations, such as companies, departments, teams, government agencies, etc. Is it still the same company after changes in employees? (Of course.) Management? (Yes.) Owners? (Maybe.) Buildings and facilities? (Yes.) Locations? (Probably.) Name? (Probably). Principal business? (Maybe.) State and country of incorporation? (Maybe.) The answers are significant to the handling of old contracts and other obligations, the determination of employee vacation and retirement benefits, etc.

And political boundaries. A data base of population statistics must have some definition of what is meant by India, Pakistan, Germany, Czechoslovakia, etc., over time. There's more involved than a change of name; the things themselves have been created, destroyed, merged, split, re-partitioned, etc. In some other data base it may have to be understood that two people born at different times in the same town might have been born in different countries.

There are some kinds of change which result in the existence of two copies of the thing, corresponding to the states before and after the change. There are several ways to deal with this situation: (1) Discard the old and let the new replace it, so that it is really treated as a change and not as a new thing; (2) Treat the old and the new as two clearly distinct things; and (3) Try to do both.

The significance of differences between copies shows up in books and other textual matter. The document you are reading now is one book. It has been and will be the "same

book" throughout a series of changes, and may even appear published in several forms with various changes in wording, punctuation, etc.

A whole spectrum of concepts. There is the "one book" containing the ideas expressed by an author, which is the same book regardless of which language it is translated into, or how it is edited, abridged, condensed, revised, etc.

Then there are "editions", which differ from each other by some arbitrary amount, due either to changes in the content or to the correction of significant amounts of error. On the other hand, some minor amount of difference (erroneous or deliberate) is permitted between reprints of a single edition.

A condensation or abridgement may be grossly different from the original, but for some purposes it is treated as being the same book.

This topic is most painfully familiar to us in relation to "versions", e.g., of such things as programs. There is some arbitrary threshold up to which minor changes can be made without creating a new version. The old copy is discarded, there may or may not be a record of the modification, and the representative (e.g., catalog entry) of the old copy now serves to represent the new copy.

Beyond a certain (arbitrary) point, we decide to keep the old and new copies as different versions. We now enter a metaphysical realm in which we manage to merge the concepts of "one" and "many", as in the expression "these several things are different versions of the same thing". In some contexts we mean to refer to all versions collectively (as in the property: this is a Fortran compiler), in some we refer to a particular copy, and in some we refer to one copy -- whichever one happens to be the "current" version.

A user who invokes the Fortran compiler several times probably believes that he is invoking the "same thing" each time even if he gets different versions. From this point of view, there should be one representative for this thing ("the current version") even though it represents different things at different times. Each version should also have its own permanent representative, and there probably should also be one representative for the collective concept of "Fortran compiler" independent of version. The representatives for the current copy and the collective concept may or may not be the same; is the property "required memory size" applicable to both?

1.4 THE MURDERER AND THE BUTLER

Combining the ideas of our last two sections: sometimes it is our perception of "how many" which changes. Sometimes two distinct entities are eventually determined to be the same one, perhaps after we have accumulated substantial amounts of information about each.

At the beginning of a mystery, we need to think of the murderer and the butler as two distinct entities, collecting information about each of them separately. After we discover that "the butler did it", have we established that they are "the same entity"? Shall we require the modelling system to collapse their two representatives into one? I don't know of any modelling system which can cope with that adequately.

1.5 CATEGORIES (WHAT IS IT?)

We have so far been focussing on the questions of "oneness" and "sameness". That is, given that you and I are pointing to some common point in space (or we think we are), and we both perceive something occupying that space (perhaps a human figure), how many "things" should that be treated as in the information system? One? Many? Part of a larger thing? Or not a thing at all?

And: do we really agree on the composition and boundary of the thing? Maybe you were pointing at a brick, and I was pointing at a wall.

And: if we point to that same point in space tomorrow (or think we are), will we agree on whether or not we are pointing at the same thing as we did today?

None of this focusses on what the thing *is*. I don't mean its properties, like is it solid, or is it red, or how much does it weigh, but what *is* it? I had to use the phrase "human figure" above because I didn't think you would follow my point if I kept using the indefinite word "thing" -- I had to convey some kind of tangible example. But that phrase is just one possible perception of the "thing" we pointed to. You might have said it was a mammal, or a man, or a solid object, or a bus driver, or your father, or a stockholder, or a customer, or ... ad nauseam.

I will refer to what a thing is -- or at least what it is described to be in the information system -- as its "category", agreeing with the usage in, e.g., [Abrial]. The same idea is also often called "type", or "entity type". Like

everything else, the treatment of categories requires a number of arbitrary decisions to be made.

There is no natural set of categories. The set of categories to be maintained in an information system must be specified for that system. In one system it might be employees and customers, in another it might be employees and dependents, or enrolled computer users, or plaintiffs and defendants, and in an integrated data base it might include all of these. A given thing (representative) might belong to many such categories.

Not only are there different kinds of categories, but categories may be defined at different levels of refinement. One application might perceive savings accounts and loan accounts as two categories, while another perceives the single category of accounts, with "savings" or "loan" being a property of each account. In another case, we might have applications dealing with furniture or trucks or machines, while another deals with capital equipment (assigning everything a unique inventory number). Thus, some categories are, by definition, subsets of others, making a member of one category automatically a member of another. Some categories overlap without being subsets. For example, the category of customers (or of plaintiffs, in a legal data base), might include some people, some corporations or other businesses, and some government agencies.

It is often a matter of choice whether a piece of information is to be treated as a category, an attribute, or a relationship. (Which raises the question of how fundamental such a distinction really is.) This corresponds to the equivalence between "that is a parent" (the entities are parents), "that person has children" (the entities are people, with the attribute of having children), and "that person is the parent of those children" (the entities are people and children, related by parentage).

It's often difficult to determine whether or not a thing belongs in a certain category. Almost all non-trivial categories have fuzzy boundaries. That is, we can usually think of some object whose membership in the category is debatable. Then either the object is arbitrarily categorized by some individual, or else there are some locally defined classification rules which probably don't match the rules used in another information system. Just as an example, consider the simple and "well understood" category of "employee". Does it include part-time employees? Contract employees? Employees of subsidiary companies? Former employees? Retired employees? Employees on leave? On military leave? Someone who has just accepted an offer? Signed a contract but not yet reported for work? Not only do the answers have to be decided according to how the company wants to treat the data, but perhaps the questions

can't even be answered consistently within the company. A person on leave may not be an employee for payroll purposes, although he is for benefits purposes. Then the notions of category and property have to be reexamined again, to arrive at a set meaningful to all users.

As another example, consider the category of "cars", and decide if the following are included: station wagons, micro-busses, ordinary busses, pickup trucks, ordinary trucks, motor homes, dune buggies, racing carts, motorcycles, etc. What about a home-made contrivance in which a short pickup truck bed is hung out of the trunk of a sedan? An old bus converted to a motor home?

As long as we are travelling, answer this question: what's the difference between a motel and a hotel? (If you have an answer, you haven't travelled much lately.)

"A more amusing example is to imagine a continuum of physical objects between some given chair and table, constructed by letting the chair back shrink while its seat expands and flattens, and its legs become higher. There will be some strange objects in this continuum which cannot clearly be assigned to either class" [Goguen]. Does the distinction between a bench and a table depend on your height?

The editor of a collection is often listed as the "author" of the book. Did he "author" anything?

The category of a thing (i.e., what it is) might be determined by its position, or environment, or use, rather than by its intrinsic form and composition. In the set of plastic letters my son plays with, there is an object which might be an "N" or a "Z", depending on how he holds it. Another one could be a "u" or an "n", and still another might be "b", "p", "d", or "q".

The purposes of the person using an object very often determine what that object is perceived to be (cf. [Stamper 77]). I can imagine the same hollow metal tube being called a pipe, an axle, a lamp pole, a clothes rack, a mop handle, a shower curtain rod, and how many more can you name? A nail driven into a wall might be designated a coat hook.

You may think you are carrying the inventory file under your arm. But the customs agent perceives a quantity of magnetic tape, and randomly snips off a sample.

Now consider some physical objects. One is a vertical rod mounted on the center of a circular stone. The second is a set of metal pointers driven around a common axis by a system of mechanical gears. The third is a marked cylinder of paraffin, with a burning cotton core. The fourth has two chambers, with a fluid flowing between them. The fifth is a

flashing digital display driven by solid state circuitry. Are these all the same kind of object? Yes -- if you happen to perceive them as clocks.

On the other hand, is a watch a clock? Of course it is -- but try asking someone if he has a clock with him.

In part, these observations illustrate the difficulty of distinguishing between the category (essence) of a thing and the uses to which it may be put (its roles).

There are also interesting questions having to do with fragments of things, and imitations. Is it still a donut after you've taken a bite out of it? Did you ever call a stuffed toy an animal?

And, like everything else, the category of an object can change with time. A dependent becomes an employee, and then a customer, and then a stockholder. A slab of marble becomes a sculpture. A piece of driftwood becomes a work of art -- just by being found and labelled! An ingot of steel becomes a machined part.

The number of entities changes, too. One ingot becomes many parts. Cutting a work of art in pieces may be vandalism -- or it may create many works of art.

Perhaps the easiest way out is to ignore the principles of continuity and conservation which we have learned since earliest childhood. It simply is no longer the same object. The sculptor does not "modify" the marble. He destroys the slab, and creates a sculpture.

The fundamental problem of this book is self describing. Just as it is difficult to partition a subject like personnel data into neat categories, so also is it difficult to partition a subject like "information" into neat categories like "categories", "entities", and "relationships". Nevertheless, in both cases, it's much harder to deal with the subject if we don't attempt some such partitioning.

For a closing amusement, do you remember "Who's On First"? Well, here's a variation:

> "Which is bigger, a baseball team or a football team?"
>
> "A football team, of course."
>
> "Why's that?"
>
> "A football team has eleven players, and a baseball team has nine."
>
> "Name a baseball team."

"The San Francisco Giants."

"How many players do they have?"

"About twenty five."

"I thought you said a baseball team has nine players."

"I guess it's twenty five."

"Any twenty five baseball players?"

"No, just the twenty five on one roster."

"If they trade a player, does that change the team?"

"Of course."

"You mean they're not the San Francisco Giants any more?"

And so on.

1.6 EXISTENCE

In a record processing system, records are created and destroyed, and we can decide with some certainty whether or not a given record exists at any moment in time. But what can we say about the existence of whatever entities may be represented by such a record?

1.6.1 HOW REAL?

It is often said that a data base models some portion of the real world. I've said so in this book.

It ain't necessarily so. The world being modelled may have no real existence.

It might be historical information (it's not real now). We can debate whether past events have any real existence in the present.

It might be falsified history (it never was real) or falsified current information (it isn't real now). Fraudulent data in welfare files: is that a model of the "real" world?

It might be planning information, about intended states of affairs (it isn't real yet).

It might be hypothetical conjectures -- "what if" speculations (which may never become real).

One might argue that such worlds have a Platonic, idealistic reality, having a real existence in the minds of men in the same way as all other concepts. But quite often the information is so complex that no one human being comprehends all of it in his mind. It is not perceived in its entirety by any agency outside of the data base itself. Or, although not overly complex, the information may simply not have reached any human mind just yet. The computer might have performed some computations to establish and record some consequence of the known facts, which no person happens to be aware of yet. It happens all the time: computers often record accounts as being overdrawn some time before any people are told about it. And even more obviously: that is precisely the point of doing hypothetical simulations by computer. The computer figures out who wins a simulated war game; in the interval between the computation and a person's reading of the output, this result is in the computer -- but what person "knows" it?

Where is the reality which the data base is modelling?

And what about fiction? The subjects of some data bases are the people, places, and events occurring in fiction (literature, mythology). This again stretches the concept of the "real world" being modelled in a data base. (Isn't fiction the opposite of reality?) But beyond that, it challenges certain premises about certain kinds of entities.

It is sometimes held that there are certain "intrinsic attributes" which all entities of a certain type must possess. For people, such attributes include birthdate, birthplace, parents, height, weight, etc. Does Hamlet have these attributes? Cities have a geographic location, an area, a population, etc. Does Camelot have these attributes?

Or shall we say instead that Hamlet is not a person, and Camelot is not a city?

Note that this situation is very different from a simple lack of information. It is not uncommon to say that we don't know a certain person's birthday, and to record it as "unknown" in the data base. That implies the possibility of eventually discovering and recording what it is. Instead, we are questioning whether such characteristics exist at all.

To conclude, if we can't assert that a data base models a

portion of reality, what shall we say that a data base does in general? It probably doesn't matter. Once again, it seems that we can go about our business quite successfully without being able to define (or know) precisely what we are doing.

If we really did want to define what a data base modelled, we'd have to start thinking in terms of mental reality rather than physical reality. Most things are in the data base because they "exist" in people's minds, without having any "objective" existence. (Which means we very much have to deal with their existing differently in different people's minds.) And, of the things in the data base which don't exist in any person's mind, whose mental reality is that? Shall we say that the computer has a mental reality of its own?

1.6.2 HOW LONG?

Some kinds of entities have a natural starting and ending, and others have an "eternal" existence; creation and destruction aren't relevant concepts for them. The latter tends to be true of what we call "concepts" -- numbers, dates, colors, distances, masses.

We could be perverse and wonder in what sort of Platonic sense such concepts have "always" existed. Did zero exist before some ancient Arab thought of it? Did gravity exist before Newton? Did the concept of television exist 50 years ago?

It doesn't really matter, for our purposes. We are not going to have to worry about creating and destroying such conceptual entities. Unless.... you are a cosmetic company, "inventing" new colors every day... or a number theorist, computing certain numbers (e.g., the primes, or perfect numbers), and adding each one to a list as you "discover" it.

There are, at the other extreme, tangible physical objects which have a well defined finite period of existence, a beginning and an end. Creation and destruction are very relevant concepts here.

But notice that I hesitate to list examples. Beginnings and endings are often processes, rather than instantaneous events. We get tied up in our definitions of what entities are in the first place. Is it the whole thing when it's partially formed? The whole abortion controversy centers on this: does a person become a person at conception, or birth, or somewhere in between? Does a car stop being a car when

it enters the junkyard? Or after it's been deformed into a solid cube?

The entity concept enters in some other ways, too. Depending on what entity categories we choose, a certain process may or may not create an entity. Hiring merely alters the attributes of a person, but it creates an employee (but be careful -- it might be a re-hire!). And, did the sculpture always exist in the marble? Recall the old vaudeville directions for sculpting an elephant: just cut away the parts which don't look like an elephant. In spite of all of this, we can entertain a notion that tangible objects have a finite existence, a beginning and an ending.

Not that we always really care. For most of our practical purposes, we prefer to treat certain objects as eternal, those whose "finite" existences appear virtually infinite: the continents, the planets, the sun, the stars. The creation and destruction of these are real only to astronomers, and to science fiction fans (real???).

But suppose that we had neatly defined tangible objects, with instantaneous beginnings and endings. Does that solve all the important problems?

We are, of course, not interested primarily in the objects themselves, but in the information we have about them. Does our handling of this information mimic the creation and destruction of such objects? Do we start having information about such objects at the instant of their creation, and stop having the information at the instant of their destruction? Of course not. We often become aware of things long after their creation (the people we deal with, the things we buy). And we're sometimes aware of them before their creation. Data are kept about children before their birth. Unborn -- and unconceived -- children are mentioned in wills. Data may kept about ordered merchandise long before manufacture begins.

And we certainly keep information about things long after they have ceased to exist.

So, does the creation and destruction of information have any direct relationship to the beginning and ending of objects? Almost never. "Create" and "destroy", when applied to information, really instruct the system to "perceive" and "forget".

Once more: we are not modelling reality, but the way information about reality is processed, by people.