

## **On the Dangers of Reading Meaning Into Surrogate Keys.**

Dr. Tom Johnston

November, 2007.

Client numbers and policy numbers like those used in the examples in this series might be assigned by business users. But it is more likely that code has generated these numbers, with "C" as a mnemonic for "client", "P" as a mnemonic for "policy", and the following three digits assigned sequentially as clients and policies are added to the database.

Assuming that this is indeed how these numbers have been created, they raise several questions. First of all, with the first character of these numbers being a mnemonic, can we still call them surrogate keys? It might appear that they are not surrogate keys, for certainly, they are being created by code which does follow a business rule, viz. the rule that client numbers will begin with "C", and policy numbers will begin with "P".

On the other hand, the three numerals following the initial letter represent, we may assume, a sequentially assigned number. And although that assignment means that the numerical part of the client and policy numbers do indicate, for any pair of clients or policies, which one became a client or policy first, we also have dates from which chronological sequence can be inferred. So we may assume here, as is in fact usually the case, that "everyone understands" that they shouldn't count on the numerical part of these numbers to indicate chronological sequence. If so, this part of the client number and the policy number is a true surrogate value, lacking business meaning.

Let's stop and be sure we understand why business users should not look to client or policy numbers as an indicator of chronological sequence of clients or policies. The basic reason is that the IT Department of the Acme Insurance Company never warranted the three numerals of those numbers as "meaning" anything. It was

generally known that they were generated sequentially, but that fact was never part of the system documentation.

What was not generally known, however, (and here we describe a situation we have come across several times in real world databases), is that policy numbers, starting about a year ago, were being re-used. Whenever an Acme policy was deleted, a six-month waiting period was observed (to allow all references to the deleted policy to be "flushed out" of the on-line database), after which the policy number was used as the unique identifier of a different policy.

As semantically risky as such reuse of identifiers obviously is, most of us who have been around for awhile have seen this kind of thing happen. In our case, Acme apparently thought that it would never have to manage more than a thousand policies. Very modest, but also a little shortsighted. And this shortsightedness is the motivation for this semantically risky reuse of policy numbers. Acme's IT department is desperately trying to put off the evil day when they will have to change from four-byte to five-byte policy numbers. And it doesn't take much reflection to realize that the changeover will be very "evil", i.e. costly and complicated, indeed.

The mere fact that such values are exposed to business users does not automatically give business meaning to those values. But when system-generated values are exposed to business users, business meaning is almost always subsequently attached to at least some of those values. For example, let's suppose that client 882's name is "Michael Smith", and that he is a very important client. We can easily imagine a conversation in which a claims processing clerk walks into her manager's office and says "Look at this. This is the third claim for 882 that has come in this week, and those claims total nearly \$125,000. It will show up on your Monday morning report, but I thought you might like a heads-up earlier than that."

In the context of this statement, it will be clear to the manager that "882" is a shorthand way of saying "client 882" or even "C882". And it is clear that the clerk is assuming that the manager knows that client 882 is a particular Michael Smith. Indeed, if the company has two or more clients of that name, then referring to the client by the identifier

the system assigned to him is advantageous because it avoids the ambiguity of whom "Michael Smith" refers to. "882" is used by the clerk and his manager to talk about a specific Michael Smith, and the *meaning* of "882", in this context, is the understanding shared by the clerk and his manager that makes it possible to use "882" to refer to that specific client.