Notes on the Open World Assumption and missingness of data

S. Henley - 27 August 2006

There are two cases of missingness of data:-

(1) Missing tuple - there is no information relating to a primary key value which is not in the relation. The primary key and its associated attributes may or may not exist (under the Closed World Assumption it does not exist).

(2) Existing tuple, but a missing attribute value. This is where a missing-data placeholder (such as the SQL 'NULL') is needed.

The missing-data placeholder can always be removed by decomposition of the relation, to convert this into case (1).

A simple example. Given date of birth - everybody has one, so Codd's 'missing and inapplicable' cannot apply. However, it may be unknown:-

1. Name and date of birth

Emp#	Name	DoB
101	A. Einstein	1 April 1965
102	J. Smith	unknown
103	I. Newton	5 Sept 1952

This should be a perfectly valid relation apart from J. Smith's unknown date of birth. In order to avoid using anything that looks like a 'NULL', however, Date would decompose this to two relations:-

2a names

Emp#	Name
101	A. Einstein
102	J. Smith
103	I. Newton

2b Dates of birth

Emp#	DoB
101	1 April 1965
103	5 Sept 1952

However, to do anything with the data (with names and dates of birth actually linked) these must be joined, and of course, unless J. Smith is to be disenfranchised simply because he doesn't know his date of birth, the 'unknown' reappears.

Looking in more detail at relation 2b, what has happened to J. Smith's date of birth ? There is no tuple for it in the relation. Under the CWA we are forced to the interpretation that because there is no tuple for Emp# 102, then for all legitimate values of DoB the corresponding predicate is false. This means that J. Smith has no date of birth. Clearly this is nonsense as it does not represent the real world.

Nevertheless, date is insistent that the CWA and its concomitent requirement for 2VL, is the only model that matches reaility. In situations where there are no missing elements or missing tuples, and all data are known, this might indeed be true: but of course in such situations the 'unknown' truth value of 3VL is not needed and the CWA is in practice identical to the OWA.

It can be shown that any occurrence of a missing data placeholder (e.g. the SQL 'NULL') in a relation can be removed by appropriate decomposition of the relation, and substituted by a corresponding 'missing tuple' in one or more of the resultant relations. If the CWA is used, such a missing tuple is necessarily interpreted as a known falsehood of the corresponding assertion - i.e. of any otherwise legitimate assertion for the given value of the primary key. The CWA is not telling us that the DoB value for J. Smith is unknown, but that he does not have a date of birth becxause of the absence of any permissible tuple for Emp# 102 in relation 2b. The CWA does not allow anything to be unknown. For that reason, in real world databases the CWA will be almost always inappropriate.