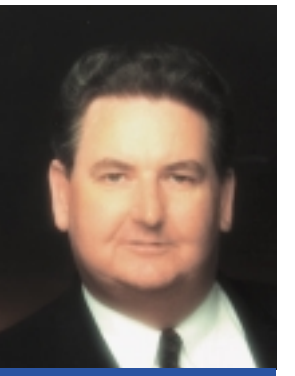
enterprise integrity



By DAVID McGOVERAN

Data Integration, Part VII

This month, we continue our simplified examination of the possible semantic relationships between a data source and a data target and the semantic transformations these imply. Understanding these can help you audit your integration efforts and plan metadata requirements.

When the source data type is a subtype of the target data type, the semantic transformation can be as straightforward as type generalization, requiring a change of units. For example, the source may be quantities of shipped pears while the target requires quantities of shipped fruit — so that the change of units is from quantities of pears to quantities of fruit.

The transformation can become more complex, however, especially if combined with consolidation. Consider what happens if some items are plums, pears, oranges, and cherries. Then the collection's semantic nature differs from that of the collection's members. Namely, we must now logically combine the semantics of the individual data types to obtain the type of the collection; the total quantity is the quantity of plums and pears and oranges and cherries. If the permissible types in the collection are always fruit and not otherwise constrained, we might give this new data type a shorthand designation of "quantity of fruit." In this case, the aggregation must be preceded by type generalization.

Sometimes the source data type and target data type aren't semantically equivalent.

A subtly more complicated example results if we permit the collection to contain bunches of grapes. Should we sum the number of grapes or the number of bunches? This is an ambiguity in the semantics of the computation and makes a difference in the semantics of the result. A similar problem occurs when we try to compute the total quantity of customers across multiple product lines or divisions. One source used households as the proper unit, another individuals, a third business entities, and a fourth shipping addresses. No rules exist within the individual business units controlling these data sources that tell how to combine the data into a semantically consistent measure of the corporation's customer base. These semantics must be defined in the interest of the consumer (or corporate management), still preserving the individual business units' semantics.

When the source data type is a supertype of the target data type, whether or not there's a semantic transformation that will suffice depends on analysis of source data values. In particular, we must be able to define a consistent decision procedure that will categorize the source data values by subtype. Only when the source data value belongs to a subtype that's semantically equivalent to (or transformable into) the target data type can we use the source data. For example, the source data type may be fruit, while the target data type is citrus fruit. We must identify, accept, and convert only those source data values that pertain to citrus fruit. The semantic transformation required is thus subtype categorization and type specialization.

Sometimes, the source data type and target data type aren't semantically equivalent, but are subtypes of a common supertype. In such cases, the semantic transformation involves type generalization to the common supertype, followed by type specialization to the target subtype. Care must be taken to ensure that both transformation portions are semantically valid. Specifically, source and target data types must not be disjointed. For example, the source data type may be various kinds of apples while the target data type may be various kinds of red fruit, with the common supertype being fruit. Some apple types are red and are red fruit; the subtypes aren't disjointed and the semantic transformation makes sense. It's equally clear that we must carefully categorize the source data, since some subtypes of fruit preclude the possibility that it's red.

Suppose the source data type and target data type have one or more subtypes in common. The only circumstance in which the source can be meaningfully used is when the source data belongs to one of those subtypes. The semantic transformation must then consist of type categorization, selection, and specialization of the source, followed by generalization to the target data type.

Next month, we'll finish semantic transformations and discuss techniques for reducing data integration's investment cost and achieving an incremental return. For now, remember that semantics pertains to more than data sharing approaches. Can you now state reasons they're crucial to all aspects of enterprise integrity? eAI

David McGoveran is president of Alternative Technologies, Inc. He has more than 20 years' experience with mission-critical applications and has authored numerous technical articles on application integration. e-Mail: mcgoveran@alternativetech.com; Website: www.alternativetech.com.