異データベース間におけるデータマッピング手法の提案（1）

— 手法確立へのアプローチ —

6R−5

石黒　正典　　坂田　哲夫　　大沼　守一
NTT情報通信網研究所

1　はじめに

1.1 背景

近年，各企業でデータベースを用いた業務のシステム化は盛んに行われているが，1企業内において，異なる業務対応にシステム化されている場合が少なくない。さらに，管理されるデータの所在まで含めた体系的なシステム化手法は確立されていない。

このため，本来システム間で共有されるべきデータが複数データベース上で重複して存在しており，データを統一的に取り扱いたいといった要求時においては，どのデータベース間でどのデータが重複しているかの明確化が必要となる。しかし，現状，全社レベルでの統一的な設計パラダイムが確立されていないことや，扱うデータのとらえ方が業務毎に異なること等により，名称の付与方法，実体・属性の表現方法，使用するコード体系等はデータベースによりまちまちである。このため，本来は同一の実体・属性・属性値を意味しているはずのデータが複数データベースにおいて存在するにもかかわらず，データベース上での表現からではそれを特定できないという問題がある（これをデータベース間の異種性の問題と呼ぶ）。

1.2 目的

複数データベース上のデータを統一的に取り扱いたいという要求に対して，スキーマ／ビューといったレベルで個々のデータベースがあたかも統合されているように利用者に見せる手法については種々の研究が行われている。[1],[2] しかし，このような統合化の手法においては，データベース間における異種性およびそれらの間の対応関係（一方のデータベースにおける要素を他方のデータベースのどの要素とどのように対応付ければよいか）は既知であることを前提としており，異種性の分析手法およびそれらの間の対応関係の導出方法については議論がなされていない。

現在，異種性の分析，対応関係の導出は人手に頼らざるを得ない状況であり，弊社においても2つのデータベース間で，実体・属性・属性値レベルでの対応関係の洗い出しに約1年近く費やした事例もある。従って，統合化の手法に加えて，効率のよい異種性の分析・対応関係の導出手法（本稿では，以降「データマッピング手法」と呼ぶ）の早期確立が望まれており，このデータマッピング手法を確立することを目的とする。

2　アプローチ

データマッピング手法確立のためのアプローチとして以下の2つが考えられる。

(1) 一方のデータベースの構成要素（例えば，テーブル／カラム）について，それと同一と見なせるものを他方のデータベースから直接導出する。

(2) 個々のデータベースで表現される現実世界の実体・属性・属性値を明確化できる統一的な記述方法を規定し，その記述方法に変換した後，同一実体，同一属性を表現する要素／要素集合をデータベース間で導出する。

(1)では，対象データベースが変わる度に分析をやりなおす必要が生じる可能性が高く，対象となるデータベースの数が少ない場合は問題ないが，数が多い場合については現実的でない。(2)は，1度分析を行ったデータベースについては，対象が変わっても再分析を実施する必要がなく，また分析を実施したデータベース間であれば，どれとでも対応関係を容易に導出することが可能となる。

我々は，扱うデータベース数も多いことから，(2)を用いてデータマッピング手法の確立を行うこととした。

3　データベース間の異種性の分類

現実世界の事物（オブジェクト）をデータベース化する際の，モデリング手法の相違，設計者による個人差，対象とする業務の特性の相違等に起因して生じる様々な異種性については，幾つか議論がなされている。[2],[3] これらは，スキーマの構成要素間に生じる異種性を中心に論じられている。我々はこれに加えて，弊社での分析事例からスキーマと同様の異種性がカラム値についても発生しうることに着目し，データベース化の対象事物を実体・属性・属性値としてとらえ，内部スキーマとしてリレーショナルモデルを前提に，表1に示すような異種性の分類を行った。

4　データマッピング手法の概要

表1からも明らかなように，異種性は「名称」と「構造」に大きく分類できる。

4.1 「名称」に対するマッピング手法

「名称」に対して規準となる記述方法（名称が表現している対象・意味を一意に識別可能な記述方法）およびそれへの変換方法を定めることにより，テーブル／カラム／カラム値(Value)における異物同名，同義の異種性を分析でき，データベース間で同一視できる要素の効果的な発見が期待できる。これについては，別稿[4]にて詳述する。

4.2 「構造」に対するマッピング手法

「名称」に対するマッピングにより，要素間の1：1の対応関係が明確化できるが，実体が複数のテーブルで表現されたり，属性が複数カラムで表現される場合があり，このような「構造」の効果的な分析手法並びに規準となる表現方法が要求される。データベースに存在する各種制約条件に着目し，その規準となる表現方法にグラフを用い，グラフ間を比較することで，データベース間の「構造」を比較する。これについては，別稿[5]にて詳述する。

An Approach for Data Mapping Method among Heterogeneous Databases

Masanori ISHIGURO, Tetsuo SAKATA, Shuichi OHNUMA
NTT Network Information Systems Laboratories

表1　データベース間における異種性の分類

項番	発生部位	分類	内　容
1	スキーマ	名称	ﾃｰﾌﾞﾙ名称間／ｶﾗﾑ名称間が異物同名の関係
2			ﾃｰﾌﾞﾙ名称間／ｶﾗﾑ名称間が同義の関係
3		構造：実体表現	同一実体が一方ではﾃｰﾌﾞﾙ，他方ではｶﾗﾑとして表現（例．出版物の著者がﾃｰﾌﾞﾙとして表現される場合と，出版物というﾃｰﾌﾞﾙの中の１つのｶﾗﾑとして表現されるようなケース）
4			同一実体の表現に使用するﾃｰﾌﾞﾙ数の相違
5			同一実体を表現するﾃｰﾌﾞﾙ間でのｶﾗﾑ数の相違
6			同一実体を表現するﾃｰﾌﾞﾙ間での従属性の相違
7			同一実体を表現するﾃｰﾌﾞﾙ間でのキーの相違
8			同一実体を複数ﾃｰﾌﾞﾙで表現する場合の参照一貫性の相違
9		構造：属性表現	同一属性の表現に使用するｶﾗﾑ数の相違
10			同一属性の表現に使用するｶﾗﾑの型・ｻｲｽﾞの相違
11			属性の隠蔽（例．社員ﾃｰﾌﾞﾙに対して，性別がｶﾗﾑとして表現される場合と男性社員／女性社員という２つのﾃｰﾌﾞﾙで表現されるようなケース）
12	カラム値(Value)	名称	値におけるｺｰﾄﾞ体系・表記方法が同一だがﾄﾞﾒｲﾝが異なる（異物同名相当）
13			値におけるｺｰﾄﾞ体系・表記方法が異なるがﾄﾞﾒｲﾝが同一（同義相当）
14		構造	値が複数ﾄﾞﾒｲﾝの組合せで構成される場合と単一ﾄﾞﾒｲﾝより構成される場合

注）カラム値(Value)　：データベース上でカラムのとる個々の値
　　ドメイン　　：現実世界において値が意味するものの集合

統合化対象となる個々のデータベースに対し，図1に示す手順に従い「名称」，「構造」の双方の観点で，設計情報の分析，規準となる表現方法での記述をそれぞれ行うことによりデータマッピングを実施する。

4.3 分析に必要な情報

分析に当たっては，データベースの設計／構築時において得られる，図2に示す各種設計情報を利用する。

5　まとめ

データマッピングを人手で行う場合，個人のスキル（データベース設計の習熟度，業務に関する知識等）に依存しており，担当により分析結果が変わるという，結果に対する一貫性が保証できないという問題もある。

本稿で述べた手法をパラダイムとして確立し，また可能な部分については自動化することにより，工数の削減と併せて，結果の一貫性の保証も可能となる。

参考文献

[1] P. Sheth and A. Larson, "Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases" *ACM Computing Surveys Vol.22, No.3, 1990.*

[2] C. Batini and M. Lenzerini, "A Comparative Analysis of Methdologies for Database Schema Integration" *ACM Computing Surveys Vol.18, No.4, 1986.*

[3] Won Kim and Jungyun Seo, "Classifying Schematic and Data Heterogeneity in Multidatabase Systems" *IEEE COMPUTER December.1991.*

[4] 大沼他，異データベース間におけるデータマッピング手法の提案（２）―データ項目間の類似性に着目したマッピング手法―，45回情処全国大会，1992

[5] 坂田他，異データベース間におけるデータマッピング手法の提案（３）―データベース内における制約条件のマッピングへの適用方法―，45回情処全国大会，1992

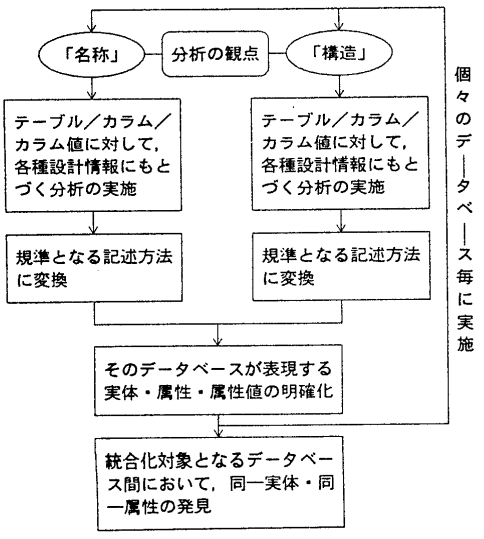

図1　データマッピングの実施手順

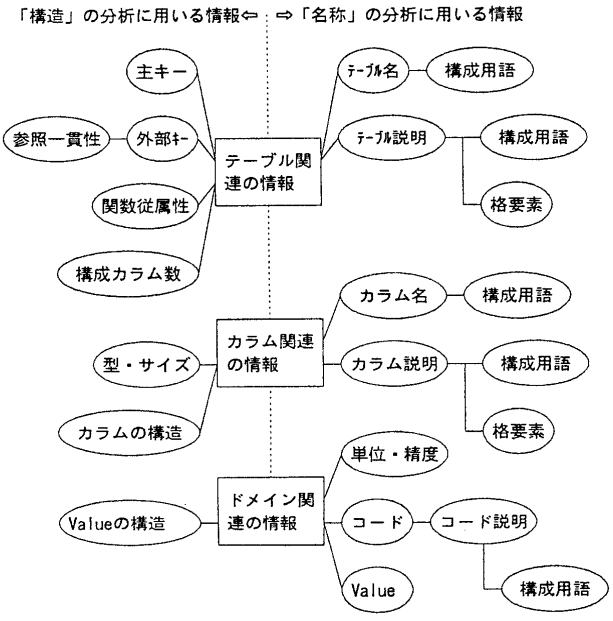

図2　分析に必要な情報の相関