体系的な DB 構築のための用語辞書を用いたデータ標準化手法

関根 純[†] 川下 満[†]
町原宏毅[†] 中川 優[†]

企業内 DB の体系的な構築のための重要な技術の一つとして，システム間で共用されるデータを識別し，そのデータに統一的なわかりやすい命名を行い，さらには，その過程で，内容，表現形式なども統一することを特徴とするデータ標準化技術が注目されている．本論文では，このデータ標準化を支援するツールの開発を通じて，わかりやすい命名のために従来から広く利用されている Durell の命名規則を計算機化する場合の問題点を明らかにすると共に，それを改善するいくつかの提案を行った．まず，データの名前を構成する用語の辞書を用いてこの命名規則を自動チェックする方法を提案した．また，この用語辞書を用いて，内容が同じで名前が異なるデータを識別可能な，効率的な類似データ分類の方法を提案した．さらに，データ標準化は，DB 設計を行う DB 管理者と企業全体のデータ管理に責任を持つデータ管理者が協調して実施する点に着目し，DB 設計との連動，プロジェクトごとの段階的なデータ標準化に有効なデータ標準化の手順を提案した．以上の提案に基づき，命名規則チェック機能，類似データ分類機能，および，用語辞書の維持管理機能などを特徴とするデータ標準化ツールを実現し，運用と定量的な評価により有効性を示した．

Dictionary-based Data Standardization Method for Database Integration

JUN SEKINE,[†] MITSURU KAWASHIMO,[†] HIROKI MACHIHARA[†] and MASARU NAKAGAWA[†]

Data standardization for identifying common data among database systems, and giving this data consistent and plain names has been one of the key technologies for constructing and integrating enterprise databases. This paper addresses new data standardization methods. Problems of implementing standard naming conventions by computers are clarified. A new method of naming data elements and automatically checking their names using a word dictionary is proposed. Also, an efficient method of clustering similar data elements using this dictionary in order to identify data elements with the same contents and different names is proposed. Then, a data standardization procedure is proposed that enables cooperation of data standardization with database design and step-by-step data standardization. A prototype of data standardization tool is developed based on these methods and the procedure. Quantitative evaluations have shown that this tool is effective in business use.

1. はじめに

近年，企業内データベース（以後，DB と略す）の体系的な構築が企業の重要な課題の一つとして認識されるようになってきた[1),5),18),19)]．このための重要な技術の一つとして，データ標準化が企業に広まってきている．データ標準化とは，DB 間で共用されるデータを識別し，DB が異なっても統一されたわかりやすい名前をそれらのデータに命名すること，および，その命名の過程でデータの内容，表現方法（使用するコード，型，桁数など）をも統一することとされている[6),8),13),15),19),23)]．

このようなデータ標準化の作業は，企業のもつデータの項目数が数千から数万，時には十万に及ぶ量になると極めて工数が掛かるため，辞書システム[1),3),4),7)-9),10),16)]が支援ツールとして利用されてきた．しかし，辞書システムが提供する機能は，主にデータの名前，内容，表現方法などの情報の管理であり，それだけではデータ標準化の作業そのものの支援を行うことはできない．

そこで，筆者らは，既開発の辞書システム[16)]と連動し，データ標準化を支援するデータ標準化ツール[8),13),15)]を開発し，社内のデータ標準化への適用を開始した．ツール開発の過程で，わかりやすいデータの命名を行うため従来から広く日米で利用されている Durell の命名規則[1)]は人間の判断を必要としそのままでは計算機化できないことがわかり，新たに，データの名前を構成する用語の辞書を利用して命名規則のチェックを自動的に行う手法，ならびに，この用語辞書を用いて，内容が同じで名前が異なるデータの識別

† NTT 情報通信網研究所
NTT Network Information Systems Laboratories

を支援する，類似データの分類手法を考案した．さらに，企業内に存在するDBシステムの数が数十以上となる場合，一時にデータ標準化を実施することは困難であることから，プロジェクト単位で段階的に，かつ，DBの設計構築と連動して，データ標準化を進める手法を提案した．最後に，データ標準化ツールの機能，および，用語辞書の構成方法等を具体化した．

2章では，データの名前，命名規則の考え方について述べる．3章では，それに基づきデータ標準化をプロジェクト単位で段階的に進める手法について述べる．4章では，用語辞書の構成について述べ，5章では，用語辞書を用いたデータ標準化の実現機能について述べる．最後に，6章で，実際のDBに適用した評価結果を示し，7章にデータ標準化ツールの実現状況，および，今後の課題等を示す．

2. データの名前と命名規則

本章では，本論文で想定するデータ標準化の対象，および，Durellの命名規則の概要とそれを計算機化する際の問題について述べる．

2.1 データ標準化の対象

アプリケーションシステムで使用するデータには様々なものがある．このうち，データ標準化の対象とするデータは，DB間で共用されることが多いデータベース，および，共用ファイル中のデータ項目とし，一時ファイルのデータ，プログラムの内部テーブル，および，内部変数などは，共用しないので，データ標準化の対象としなかった．

このデータ項目に付与される名前が備えるべき条件には，データを共用するという観点から挙げると，次のものがある[19]．

一意性：同じ内容を表すデータ項目は同じ名前，異なる内容を表すデータ項目は異なる名前を持つこと．ここで同じ内容を表すとは，現実世界の同じ対象に対して，同じ概念を表すと人間が認識することを指す．以後，本論文では，この考え方に従う．

一貫性：企業全体として同じ名前の付与方法を用いていること．

識別性：名前から内容を容易に理解できること．

安定性：アプリケーションの実現法の変更に強いこと．

操作性：アプリケーション構築時にコーディングが容易であること．

しかし，上に挙げた要件には，識別性を重視した長い名前を付与すると，コーディング時の操作性が悪くなるなど背反のものがあるため，単一の名前を用いて上記の要求をすべて実現することは難しい．そこで，実際には，目的に応じて複数の名前を使い分けている．本論文では，名前を次の3種類に分類した．

(1)日本語名：エンドユーザ，および，DB設計を担当するデータベース管理者が内容を識別するための日本語の名前．識別性，安定性を重視した名前である．

〈例 2.1〉 日本語名：電話番号

(2)定義名：特定のソフトウェア製品（例えば，DBMS）の利用の際にプログラマが用いる名前であり，その製品の制約（文字数の制約，予約語は使用不可など）に従う必要がある．コーディングの容易さ，製品の制約から，アルファベットが使用されることが多い．操作性を重視した名前である．

〈例 2.2〉 定義名：dnw_bng

(3)ディクショナリ名：企業全体のデータ管理に責任を持つデータ管理者が用いるのに適した，一意性，一貫性を重視した日本語の名前．ディクショナリ名は，文献17),19),22)に記述が見られるように，特定の概念モデルを前提に，実体，実体間の関係，役割，ドメインなどの概念の組み合わせにより厳密に構成されるものである．

〈例 2.3〉 ディクショナリ名：
顧客．主キー．電話番号．文字列

このディクショナリ名を用いるとディクショナリにおけるデータの管理は容易になるが，そのためには，名前の付与以前に，企業が保有するデータの全体像を表す概念モデルが明確になっている必要がある．しかし本論文では，必ずしもこれが明確でない段階でも利用可能としたい，また，データベース管理者やエンドユーザ等にもわかりやすい自然な名前としたいとの要求から，(1)，(2)の2種類の名前をデータ標準化の対象とした．

また，COBOLに見られるような階層的なデータの定義法は，利用者やデータベース管理者の理解を助けることから，データ項目として，次の例に示すような階層的な名前の構成（構造体）を許すことにした．

〈例 2.4〉 割賦
支払回数
支払残回数

なお，データ標準化を達成した段階であれば，これ

ら(1),(2)の二つの名前をディクショナリで管理することで十分と考えられるが,本論文では,一気にデータ標準化を行うことは困難な状況を想定している.そのため,特定のシステムでのみ通用する,標準化が済んでいないローカルな名前,および,企業内で通用する標準化済みのグローバルな名前の両者が共存する.そこで,データ項目の名前は,標準,非標準あわせて4種類管理することにした.

2.2 データ標準化とは

2.1節の検討を踏まえ,データ標準化とは,次の要件を満たすようにデータ項目を整備することであると定義する.

要件1:日本語名は,業務やDBに必ずしも精通していないエンドユーザにもわかりやすいこと.

要件2:定義域が同じで,かつ同じ内容を表すデータ項目は,日本語名が同一であること.逆に,これが満たされないデータ項目の日本語名は異なること.ここで,定義域が同じとは,顧客電話番号と従業員電話番号の例に示すように,取り得る値(電話番号)が同じであることを意味するものとする.なお,定義名の標準化は特定の製品の制約の影響を受け困難であるため対象外とした.以後断らない限り,データ項目名とは,日本語名のこととする.

要件3:定義域が同じデータ項目の表現形式(型,桁数,コード値)は同一であること.

要件1を具体化し命名規則としたものがDurellの命名規則[1),23)]であり,これはおよそ,次のような規則である(**図1**).

(1)データ項目名は,複数の用語から構成される.

(2)データ項目名の右端には,そのデータ項目がどのような値を持つかを明確にする,「番号」,「名」などの用語を必須とする.これを区分語と呼ぶ.

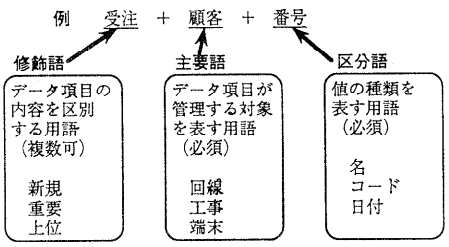


図1 Durellのデータ項目命名規則
Fig. 1 Durell's naming conventions for data elements.

(3)区分語の左には,どのような実体に対するデータ項目であるかを明らかにする,「契約」,「社員」などの用語を必須とする.これを主要語と呼ぶ.

(4)主要語の左には,0または1つ以上の,意味を補う用語を付与可能である.これを修飾語と呼ぶ.

しかし,この命名規則は,どの用語が主要語であり区分語であるかがわからなければ,計算機化は難しい.そこで,本論文では,この用語と用語の種別(区分語,主要語,および,修飾語となるか否かを表す分類)を用語辞書に持つことにより,命名規則を自動的に判定可能とする手法を考案した(5.1節).さらに,要件2,3に違反したデータ項目を発見するために,この用語辞書を用いて,同じ内容で異なる名前を持つデータ項目の識別を支援する,データ項目名の分類手法を考案した(5.2節).文献1)には,このうち,命名規則のチェックを行うツールの記述が見られるが,定義名と日本語名を区別していない点で本論文と前提が異なること,および,米国で使われているツールであり,英語の品詞を考慮した作りとなっているため日本語名に適用できないことなどの問題がある.

さらに,データ標準化の副次的な効用として,従来より,日本語名を構成する用語ごとにアルファベットの省略語を割当て,この省略語の組み合わせにより定義名を自動生成する技術が提案[19)]されているが,この提案には,自動生成した定義名がDBMSに存在する定義名の長さ制限に違反するという問題,および,異なる日本語名から同じ定義名が生成されるという一意性の問題があった.そこで,上に述べた用語辞書に複合語の考え方を追加することにより,この問題を解決する手法を考案した(5.3節).

次の3章では,まず,このようなデータ標準化技術の前提となるデータ標準化の手順について述べる.

3. データ標準化の実施環境と手順

3.1 データ標準化の実施環境

本論文では,データ標準化を次のような環境で行うことを前提とした.

(1)データ標準化の稼働は極めて大きいことから,DBの企業内全体計画に責任を持つ管理部門のデータ管理者が指導し,実際の作業は,DB設計を担当するプロジェクトのデータベース管理者が分担して実施する.

(2)データ標準化,および,DB設計は並行して実施する.

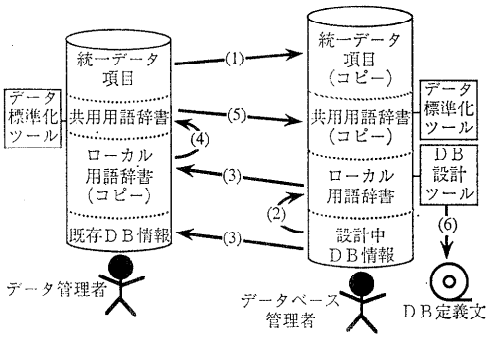


図 2 データ標準化の手順
Fig. 2 Typical data standardization procedure.

(3)データベース管理者，および，データ管理者は地理的に離れているとの想定のもとに，データ標準化ツールは分散システム構成とする．

具体的には，既に開発済みの DB 情報や，企業全体で使用する用語辞書，統一コードなどを管理するホストシステム，および，特定のプロジェクトに固有の開発中の DB 情報を管理するローカルシステムからなる分散システム形態を採用し，ホストシステムには，データ標準化ツール，ローカルシステムには，データ標準化ツール，および，DB 設計ツールの両者が存在する構成とした（図 2）．

3.2 データ標準化の手順

典型的な実施例として，プロジェクト単位に段階的にデータ標準化を行う場合の，データ標準化ツールの利用手順を次に示す．

3.2.1 新規 DB の構築

新規に DB を構築する場合，DB の設計段階から計画的に，2 章で述べた要件を満たすように全データ項目を作成できる．これは次の手順による（図 2 の(1)から(6)）．

(1)データベース管理者は，まず，データ管理者が提供する統一データ項目が利用できるならそれを用いて，利用できないなら独自にデータ項目を作成し，自プロジェクトの DB の概念設計を行う．ここで，統一データ項目とは，複数の DB で共用される可能性が高いデータ項目を，データ管理者が社内標準と定め，日本語名，型，桁数などを統一して管理しているものを指す．

(2)次に，データベース管理者は，データ項目名を命名規則に従ってチェックし，2 章に示した要件 1 を満たしていることを確認する（5.1 節）．データ管理者が配付した共通の用語辞書にない新たな用語が発生したら，それをプロジェクト固有のローカルな用語辞書に入れ，新たな共通の用語の候補として管理する．

(3)データ管理者は，データ項目の設計結果，および，用語の候補をデータベース管理者より受け取り，前者については，2 章の要件 2，3 の観点で問題がないかチェックを行う．このチェックを行うため，データ管理者は，保有する全 DB のデータ項目と突き合わせて，5.2 節に述べる類似データ項目の分類を行い，似た名前のデータ項目を相互に比較し，同じ内容のデータ項目か否かの判断を行う．同じ内容のデータ項目であるなら，同じ日本語名，型，桁数とすべきことを，データベース管理者に通知する．

(4)用語の候補については，データ管理者は，a．共通の用語辞書にそれと類似の用語がないか，b．共通の用語として利用するのに価する企業内で共通な用語か，などの観点で分析をし，問題がなければ用語を共通の用語辞書にマージする．

(5)新しい共通の用語辞書は，データベース管理者に再配付される．データベース管理者は，データ管理者からの通知に基づきデータ項目名の修正を行った後，新たに配付された共通の用語辞書と重複するローカルな用語辞書の用語を削除する．さらに，データ項目の日本語名が確定したら，その日本語名を元にデータ項目の定義名を作成する（5.3 節）．

(6)以上の手順の後，DB の論理設計，物理設計を行い，最後に DB を生成するための定義文を作成する．

3.2.2 既存 DB の更改

運用中のシステムにおけるデータ項目名の統一は，稼働が大きい割にリスクが大きく，直接的な効果が少ないことから困難である．そこで，システム更改の時期に合わせて変更を行うことになる．それまでの間に，データベース管理者，および，データ管理者は次の事前準備作業を行っておく．

(1)既存の DB に関する情報をディクショナリに入力する．筆者らは，これらの情報が既存 DB 中のスキーマよりもむしろ，ワープロで作成された文書ファイルに存在することに着目し，そのようなワープロ文書から，自動的に情報を抽出するツールを作成して効率化を図っている[11]．

(2)新規 DB の構築の場合と同様にデータ項目を他の DB と比較して分析し，その結果，理想とすべきデータ項目の標準日本語名を設定する．この時点で，データ項目は，実際の DB で用いられている非標準な名前と，標準として定めたデータ項目名の両者を対応付けて持つことになる．

以上の作業の後，システム更改時に実際の DB で用いられる非標準な名前を標準の名前に変更する．

4. 用語辞書の構成

以上に述べたデータ標準化の作業のために必要となるディクショナリの構造は，IRDS 上に構築されている（**図3**）[16]．このうち用語辞書は次のような情報を持っている．

(1)用　　語

それ一語で意味を持つ最小の単位．管理対象の全 DB の全データ項目の日本語名に現れる用語を管理する．用語は，一般的には品詞と合致する（詳細は文献13)参照)．

(2)標準用語

用語の持つ表記法の違いを吸収し統一した用語．例えば，下記の例 4.1 の場合，三つの用語はすべて右の標準用語「ユーザ」に統一される．

〈例 4.1〉 用　　語　　標準用語

なお，どの表記を標準用語とするかについては，標準化作業で広く使われている文献21)の規則を採用した．標準用語には，次の情報が含まれている．

a．標準用語種別

その標準用語が区分語となるのか主要語となるのか，修飾語となるのかを表す種別．実際には，標準用語の種別は一つに固定されるわけではなく，状況に応じて変わりうる．例えば，データ項目「電話番号」では標準用語「番号」は区分語であるが，データ項目「電話番号払い出し日付」では，「番号」は修飾語となる．商用システムにおけるおよそ2万1千のデータ項目の標準用語を用いて検証した結果，標準用語は次の3パターンに分類できることがわかった．そのパターン名を用語辞書に登録している．

パターン1：区分語，主要語，修飾語になりうる標準用語（番号，名，数など）

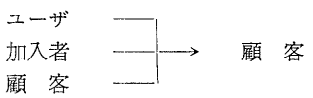
パターン2：主要語，修飾語になりうる標準用語（端末，回線，顧客など）

パターン3：修飾語にしかならない標準用語（新，重要，今期など）

b．代表標準用語名

標準用語としては異なるが，類似の内容を持つものをグループ化して，そのグループを代表する標準用語を決めている．これを代表標準用語と呼ぶ．この情報は，類似データ項目の分類時に使用される．例えば，下記の例 4.2 では，左に示した標準用語はすべて，同じ代表標準用語「顧客」を持つ．

〈例 4.2〉 標準用語　　代表標準用語



c．略　　称

企業内でのみ使われる特殊な略称については，それがどのような標準用語の組み合わせの短縮形であるかを記述する．この情報は，類似データ項目の分類時に使用され，短縮前に同じ標準用語を持つデータ項目は，類似，あるいは，同一と見なされる．

〈例 4.3〉 電番→電話＋番号

d．標準用語定義名

標準用語に割り当てたアルファベットの名前．本論文では，この名前の組み合わせでデータ項目の定義名を自動生成する．このアルファベット名自体も自動生成される（5.3節）．

e．禁止語フラグ

標準用語は，用語の表記の違いを吸収しただけのものであるから，次に示すいくつかの理由から使うのが望ましくない標準用語も含まれている．そこで，このフラグにより，当該の標準用語が禁止語であることを宣言し，5.1節の機能を用いてエラーとしてチェックアウトする．禁止語の例としては，

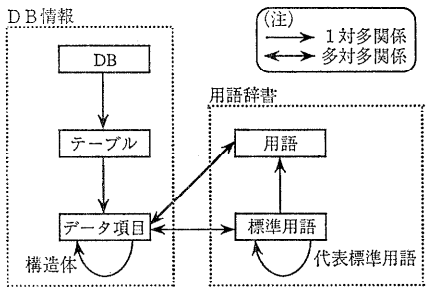


図 3　情報資源辞書の構造
Fig. 3　Structure of information resource dictionary.

「顧客データ」に示す「データ」,「受注情報」に示す「情報」のように，その標準用語を用いても語感を整えるだけで情報量が増えない標準用語，あるいは，企業の特定の人にしか理解できないローカルな標準用語，あるいは，順番を表すために用いる単独のアルファベットや「甲乙丙丁」などの標準用語がある．順番を表す標準用語を禁止語にしている理由は，これが第1正規化（繰り返し項目を持たない）に違反しているデータ項目を示す可能性があり注意を促すため，および，やむをえずデータ項目に繰り返しを許す場合でも統一のため繰り返しを表すのに数字を用いることを強制するためである．

データ標準化ツールの機能は，この用語辞書を用いて5章のように実現されている．

5. データ標準化ツールの機能

5.1 データ項目の命名チェック機能

データ項目名のチェックは，Durell の命名規則に合致しているか，禁止語を用いていないか，および，用語辞書に登録されている用語を用いているかの三つの観点から行うことにした．なお，これらのチェックのためには用語辞書が必要であるが，データ項目名を構成する用語がすべて用語辞書に含まれている保証はないため，命名規則によるデータ項目のチェック時に，新たに発見した用語の候補をローカルな用語辞書に蓄積する必要がある．これは，次のアルゴリズムに基づき行われる（**図4**）．

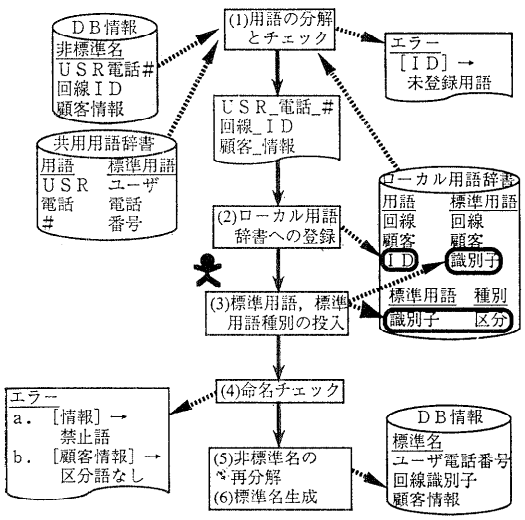


図4 データ項目名チェックの手順
Fig. 4 Data element name checking procedure.

(1)データベース管理者は，あらかじめ，データ項目名を非標準の日本語名としてDB情報に投入しておく．ツールは，これらのデータ項目名を，既にある共用の用語辞書，および，ローカルの用語辞書を用いて用語に分解し，用語の区切りにアンダスコアを挿入する．同じ文字から始まる用語がある場合には，文字数の多い用語から優先的にマッチングを行う．これは自然言語処理における最長一致法に相当する．両者の用語辞書にない用語をツールに認識させてローカルな用語辞書に投入するため，投入者があらかじめ用語の区切りにアンダスコアを入れておくと，用語辞書を用いることなく，その区切りを信じて用語への分解を行う．ただし，用語分解の結果得られる用語が，用語辞書に登録されていなければエラーであることをレポートする．

〈例 5.1〉 USR 電話#→USR_電話_#

(2)共用，および，ローカルの用語辞書に登録されていない用語を，ローカルな用語辞書に自動的に登録する．

(3)ローカルの用語辞書に蓄えた用語に対応する標準用語，標準用語種別などは人間が新たに投入する．これを行った後，再度，(1)からの手順を繰り返すと，最終的には，(1)で生じたエラーは消える．

(4)次に，以下の方法で命名規則に合致するかをチェックする．

a．その用語の標準用語を調べる．標準用語が略称であるなら，略称を標準用語に展開する．標準用語が禁止語であれば，エラーとする．

b．各標準用語の用語種別を調べ，それに基づき，Durellの命名規則に合致するかをチェックし，違反の場合エラーとする．

c．データ項目が構造体の場合には，最上位から最下位までのデータ項目名を接続したものについて，a.，b. を適用する．

(5)非標準のデータ項目名を，共用の用語辞書，および完成したローカルな用語辞書を用いて用語に再分解し，用語の区切りにアンダスコアを挿入する．この操作が必要な理由は，(1)において投入者がアンダスコアを挿入することにより指定した用語の区切りは信頼性が低く，間違いが多いことによる．

(6)非標準のデータ項目名に含まれる用語を標準用語

```
<禁止語チェック>
リース情報　　：禁止語［情報］が使われています.
収納状況履歴A：禁止語［A］が使われています.
氏名･住所表示 ：禁止語［･］が使われています.
一時課金データ：禁止語［データ］が使われています.
・・・・・・・・・・・・・・・・・・・・

<命名チェック>
前月料金：［前月］は修飾語で，規則に合いません.
　　主要語であること.
連絡先　：［先］は修飾語で，規則に合いません.
　　区分語であること.
竣工サービスオーダ：［サービスオーダ］は主要語で
　　規則に合いません．区分語であること.
年月　　：［年月］の左に主要語がありません.
・・・・・・・・・・・・・・・・・・・・
```

図 5 データ項目名チェックの出力結果
Fig. 5 Outputs of data element name checking.

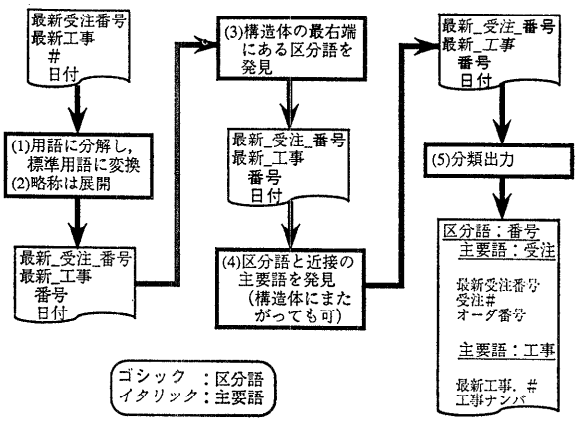


図 6 類似データ項目分類の手順
Fig. 6 Procedure of clustering similar data elements.

に自動変換して，標準のデータ項目名を生成する．生成した標準のデータ項目名に問題がある場合，それをデータベース管理者が修正することは可能である．

〈例 5.2〉 USR_電話_#
→ユーザ_電話_番号

以上の手順により，データ項目の日本語名をチェックすると同時に，用語辞書の構築，標準の日本語名の生成を可能にしている．

なお，データ項目名のチェック方法としては，大量のデータ項目をチェックせずに一括投入した後，一括チェックを行う方法と，データ項目を1件投入するごとにチェックを行う方法を用意しているが，投入者と分析者の作業分担ができる点で前者の方が利便性が高く，利用頻度が高い．**図5**にチェック結果のレポート出力例を示す．

5.2 類似データ項目分類機能

本機能の目的は，同一の定義域，あるいはさらに，同じ内容を表すデータ項目の発見を支援することである．データ標準化が行われていないDB間では，同じ内容でもデータ項目名が微妙に異なることが頻繁に見受けられる．そこで，似た名前のデータ項目をツールが発見してデータ管理者に示せれば，利便性は高い．しかし，データベース検索やフルテキスト検索[2),12),14),20)]でよく用いられる，用語ごとに類似の用語を見つけ，それを元にマッチする用語の数が多い複合語を見つける，といった単純な類似検索の方法では，明らかに定義域が異なる「最新_受注_番号」と「最新_工事_番号」も二つの用語が一致するため類似と判定され，精度が良くならない．

そこで，筆者らは，実データの分析による定性的な検証から，同一の定義域を持つデータ項目は類似の区分語，主要語を持つ可能性が高いことに着目し，区分語と主要語を発見し，それぞれを大分類，小分類とするデータ項目の分類を行うことで類似データ項目の分類の精度向上を可能にした．次にアルゴリズムの概要を示す（**図6**）．

(1)データ項目名を用語に分解し，それぞれを用語辞書を用いて標準用語に変換する．
(2)略称がある場合，略称を標準用語に展開する．
(3)データ項目名を構成する標準用語を末尾から順に左に探し，区分語になりうる標準用語を探す．データ項目名が構造体を構成する場合，下位のデータ項目から上位のデータ項目に区分語を見つかるまで探す．
(4)見つかった区分語から，さらに左方に，また，上位のデータ項目に主要語を探す．
(5)見つかった区分語の代表標準用語を求め，それを用いてデータ項目名を大分類する．さらに，大分類した結果のデータ項目名を主要語で小分類する．この結果，類似のデータ項目が一つの小分類の中に集約される．

本方式では，あるデータ項目に類似のデータ項目は，そのデータ項目が存在する小分類の中に絞り込まれており，データ管理者は，この小分類に含まれるデータ項目を網羅的に調査することにより，最終的に，同一の定義域，あるいは，同じ内容を表すデータ項目を決定できる．

このようなデータ項目の分類をDBごとにローカルシステムで行い，DBごとの類似データ項目分類の中間結果はホストシステムに送付して保管することにした．この中間結果のマージ処理を行うことにより，任

意のDB間で類似データ項目を比較できる，分散システム構成での効率的な類似データ項目分類方式となっている．

5.3 定義名自動生成機能

データベース管理者にとって，データ項目の定義名の生成は本質的な作業ではないが，プログラム設計，DB 設計が終了したコーディング以降に必要となり，現状での作業量は多い．この定義名の生成を自動化できればその利便性は高い．データ項目の定義名を生成するアルゴリズムの概要は文献1),19)に見られ，それは次のアルゴリズムに依っている．

(1)データ項目の定義名は，その日本語名を分解した結果えられる標準用語の定義名をアンダスコアでつないだものとする．

〈例 5.3〉 電話番号→電話＋番号
→dnw_bng

(2)標準用語の定義名は，標準用語の読みをローマ字化し，その文字数を減らすため，母音を後ろから除いていったものとする．ただし，異なる標準用語で同じ定義名が出現すると，定義名から日本語名を一意に識別することが困難になるため，母音の削除の方法を調整して一意となるようにする．

〈例 5.4〉 電話→denwa→dnw

しかし，この方法には，次の例5.5に示すように標準用語に同音異義語があるとその定義名を一意とすることが困難であること，および，実際に適用するとデータ項目の定義名が著しく長くなり，使用する DBMS の文字数制限を越える場合が多く現われるという問題があった．そこで本論文ではこれを次の点で改良したアルゴリズムを具体化することにより解決した．

〈例 5.5〉 間，管，巻→kan

(1)異なる標準用語で同じ定義名が出現した場合には，末尾に数字を付けて一意となるようにする．

〈例 5.6〉 管→kan
間→kan 2

本論文で想定するデータ標準化ツールは一つのホストシステムと複数のローカルシステムからなる分散構成であり，定義名の生成はローカルシステムで自律的に行われるため，上記の数字を一意に払い出すことは難しい．そこで，標準用語の候補をホストシステムにおいて共用の用語辞書に統合する際に，標準用語の定義名が一意か否かを判定し，定義名が重複する場合にはホストシステムで修正し，ローカルシステムに反映する方式としている．

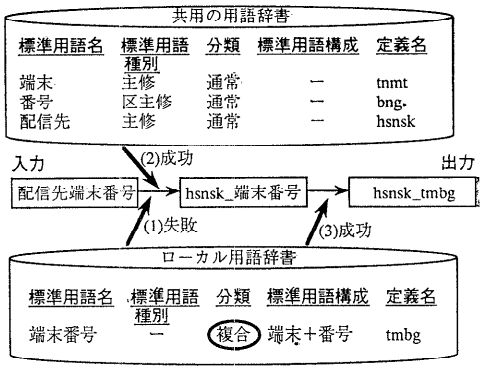


図 7 複合語を用いた定義名の自動生成
Fig. 7 Automatic creation of definitive name using compound words.

(2)用語辞書に複数の標準用語から構成される複合語を登録可能とし，この複合語ごとに定義名を設定可能とした．また，複合語がどのような標準用語から構成されるかを略称と同様に記述することにした．これにより，出現頻度が多い複合語の定義名の長さを短くすることが可能となった．次の例では，「端末番号」という複合語を登録することにより定義名の短縮を行っている（図7）．

〈例 5.7〉 配信先端末番号→
配信先＋端末＋番号→
hsnsk＋tnmt＋bng→hsnsk_tmbg

この方式に従うと，ある DB では，「端末番号」の定義名は tnmt_bng，別の DB では，tmbg と異なることになるが，本論文の場合，データ項目の定義名は共用する価値が低いとの判断からデータ標準化を行わないとの前提に立っているため，不都合は少ない．そこで，複合語はローカルの用語辞書での管理とし，共用の用語辞書への統合の対象外としている．

以上の複合語を考慮した分解を可能とするために，用語辞書を用いたデータ項目の標準用語への分解の時，共用の用語辞書より先に，複合語を含むローカルの用語辞書を用いて標準用語へのマッチングを行う方式を用いている．

6. 評 価

まず，本論文で提案した命名規則チェック適用の効果を示すため，既存の 14 DB, 12,280 データ項目，および，データ標準化ツールを用いて新規に DB を設

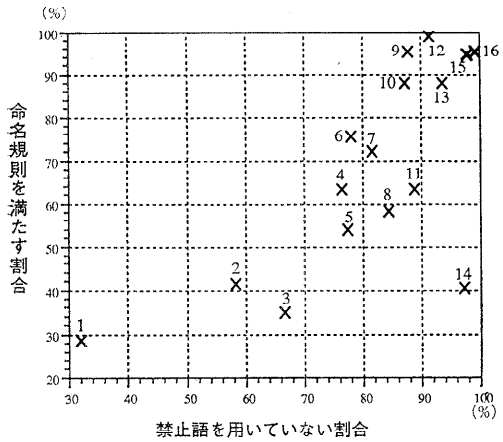


図 8　データ項目名チェックの結果
Fig. 8　Results of data element name checking.

計した結果の 2 DB, 1,326 データ項目が，どの程度 Durell の命名規則に従っているか，および，禁止語を用いているかの割合を DB ごとに評価し，プロットした（**図 8**）．その評価によると，既存の DB の場合，命名規則にデータ項目が違反している割合は，41％であるのに対し，データ標準化ツールを用いて設計した DB の場合，違反している割合は 5％と改善されていることがわかる．これが，0％とならない理由は，社内で慣用として使っており変更できない名前があり，それが命名規則に違反していることによる．一方，禁止語については，既存の DB の場合，違反している割合が 17%，データ標準化ツールを用いて設計した DB の場合，1％弱となっており，同様に改善されていることがわかる．

このような分析の過程で得られた用語の数は，共用の用語辞書においては，用語数 1,692 項目，標準用語数 1,653 項目となり，共用の用語辞書にまだ統合していないローカルの用語辞書においては，用語数 2,005 項目，標準用語数 1,800 項目となった．さらに，同一の業務分野であれば，限られた数の標準用語でデータ項目名を構成できるのではないかという予想の元に，同一の業務分野を支援する 10 DB（4,545 データ項目，715 標準用語）を上記の DB の中から選択し，データ項目数の伸びに対する標準用語数の伸びを求めた（**図 9**）．図 9 からわかるように，同一の業務分野であれば，データ項目数に対する標準用語数の伸びは漸減する．定量的に見ると，データ項目の初期投入段階では，1 件のデータ項目の追加に伴い平均 1.45 の

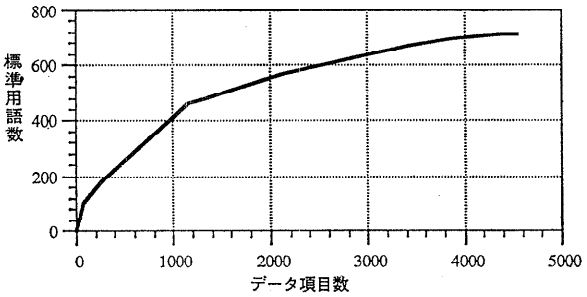


図 9　データ項目数に対する標準用語数の変化
Fig. 9　Variation of number of standardized words with number of data elements.

標準用語の追加が発生するのに対し，データ項目の登録数が 4,500 を越える所では，1 件のデータ項目の追加に伴い，平均 0.08 の標準用語しか追加されないことがわかった．すなわち，用語辞書を拡充してゆけば，同じ業務領域については，次第に標準用語の追加が不要になり，用語辞書の維持管理のためのデータベース管理者，データ管理者の負荷はほとんどなくなることがわかる．

次に，データ項目の命名規則に従うことにより，データ項目名がどのように変化するかを把握するため，1 データ項目に含まれる用語の数を調査した．その結果，命名規則を適用しない DB においては，用語の数は平均 2.49 であるのに対し，命名規則を適用した DB では，用語の数は平均 5.05 であることがわかった．すなわち，命名規則を適用してわかりやすいデータ項目名とすることにより，データ項目を構成する用語の数が倍増することがわかる．

次に，類似データ項目分類の評価結果を示す．上記とは異なるシステム（70 DB, 7,869 データ項目）に類似データ項目分類を適用し区分語，主要語で分類したところ，平均 35 個のデータ項目からなる 227 個のグループに分類できた．このことから，本方式によれば，あるデータ項目に類似のデータ項目は，平均 35 個に絞り込めたことがわかる．一方，従来の類似データ項目検索方式では，類似のデータ項目が平均 330 個現われた．すなわち，データ管理者が調査すべきデータ項目数が本方式の 9 倍にも及ぶことがわかり，本方式の有効性が確認できた．また，70 の DB における 7,869 のデータ項目が実際には 1,198 の異なるデータ項目に集約できることがわかり，DB 間で大量のデータ項目の重複があることも確認できた．

最後に，本論文で提案したデータ項目定義名の生成方式の評価結果を示す．新規に設計した 2 DB, 1,326

データ項目に本生成方式を適用し，従来方式と比較評価した．これらの DB では，DBMS 製品の制約として定義名の長さは 18 文字以下という制限がある．従来の生成方式では，208 データ項目（16%）がこの制限を越え，最長の定義名の長さは 45 文字，平均の定義名の長さは 13.6 文字であった．本方式を用いて，用語辞書に 228 の複合語を加えることにより，定義名の長さをすべて 18 文字以下に抑え，平均の定義名の長さを 10.7 文字とすることができた．

7. おわりに

本論文では，DB 設計と連動し，プロジェクトごとに段階的にデータ標準化が可能なデータ標準化ツールの実現方法について明らかにした．

まず，用語の種別を持つ用語辞書を利用することにより，命名規則を自動的にチェックする手法を提案した．また，用語の種別と類似性を元に，定義域が同じ，あるいは，同じ内容を表すデータ項目を効率よく探すことが可能な分類手法を考案した．次に，データ項目の日本語名から定義名を生成する従来の手法に生成結果の定義名が長くなるなどの問題があることを示し，改善案を提案した．

さらに，プロジェクトのデータベース管理者と管理部門のデータ管理者が協調してデータ標準化を行う手法を提案し，これが DB 設計との連動，および，プロジェクトごとの段階的なデータ標準化に有効であることを示した．最後にこれらが実際に有効であることを，実システムでの命名規則の評価により定量的に示した．この評価により，同じ業務領域であれば，データ項目数が増えても新たな標準用語の発生は頭打ちとなり，データ項目名は限られた標準用語の組み合わせで構成できることも確認できた．また事業への適用により，用語辞書の維持管理が DB 設計の現場でも十分可能であることがわかった．

データ標準化ツール，および，DB設計ツールは，設計情報を集中管理する情報資源辞書システム(IRDS)[16]を核としたDB設計支援システム(DBprompt)[11]を構成し，SUN ワークステーション上で現在稼働している．このうち，データ標準化ツール (DBprompt/NAME) はC言語で 90 キロステップの規模を持つ．IRDS を実現する DBMS としてリレーショナルデータベース管理システム INFORMIX を利用した．DBのサイズは，およそ 50 MB ある．

今後の課題は，データ標準化ツールと CASE ツールとの連動，および，標準用語の種別の設定方法を定式化（一部，文献8)，13)に記述済み）することである．

謝辞 実際にデータ標準化ツールを使用し，様々な情報を提供して頂いた NTT 通信ソフトウェア本部，大沼主幹技師，森野主任技師，黒川社員，および，NTT ネットワーク高度化推進本部の林担当部長，宮崎主査に感謝いたします．データ標準化ツールの研究と開発にあたってご指導を頂いた NTT 情報通信網研究所データベース研究部，伊土部長，石垣主幹研究員，田中主幹研究員に感謝いたします．本研究のきっかけを与えて頂いた，データ標準化に興味のあるユーザの団体である IRM 研究会の諸氏に感謝いたします．

参考文献

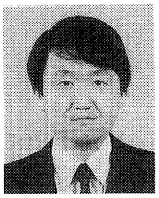
1) Durell, R. W.: データ資源管理，日経マグロウヒル (1987).
2) Faloutsos, C. and Chan, R.: Fast Text Access Method for Optical and Large Magnetic Disks; Designs and Performance, *Proc. Int. Conf. Very Large Data Bases*, pp. 280-293 (1988).
3) Hsu, C., Bouziane, M., Ratter, L. and Yee, L.: Information Resource Management in Heterogeneous, Distributed Environments; A Metadatabase Approach, *IEEE Trans. Softw. Eng.*, Vol. SE-17, No. 6, pp. 604-625 (1991).
4) *Information Resource Dictionary System (IRDS) Services Interface*, ISO DIS 10728, ISO/IEC JTC 1/SC 21 (1991).
5) Martin, J.: *Strategic Data-Planning Methodologies*, Prentice-Hall Inc. (1982).
6) Newton, J. J.: *Guide on Data Entity Naming Conventions*, NBS Special Publication, 500-149 (1987).
7) Sagawa, J. M.: Repository Manager Technology, *IBM Syst. J.*, Vol. 29, No. 2, pp. 209-226 (1990).
8) Sekine, J., Nakagawa, M., Kimoto, H. and Kurokawa, K.: A Standard Naming Method of Data Elements Using a Semantic Dictionary, *3rd Int. Conf. Database and Expert Systems Applications*, pp. 167-172 (1992).
9) The BACHMAN/DBA DB 2; A Practical Approach to DB 2 Database Design, *InfoDB*, Vol. 3, No. 2, pp. 24-31 (1988).
10) 岩崎一正，穂鷹良介：ISO IRDS の実装と機能の吟味，情報処理学会データベースシステム研究会，87-2, pp. 9-16 (1992).
11) 大久保成隆，町原宏毅，関根 純，中川 優：DB 設計支援ツール DBprompt のアーキテクチ

ャ，第44回情報処理学会全国大会論文集，4-237-238 (1992).
12) 菊池忠一：日本語文書用高速全文検索の一手法，電子情報通信学会論文誌，Vol. J 75-D-I, No. 9, pp. 836-846 (1992).
13) 黒川 清，中川 優，関根 純：データ標準化ツール (DBprompt/NAME) における複合語解析を用いた用語辞書構築方法，第44回情報処理学会全国大会論文集，4-239-240 (1992).
14) 小林清浩：入力文字列に対するファイル内類似文字列の探索法について，電子情報通信学会論文誌，Vol. J 74-D-I, No. 1, pp. 39-49 (1991).
15) 関根 純，川下 満，鈴木健司：ネーミング手法と支援ツール，信学技報，DE 89-4 (1989).
16) 関根 純，川下 満，中川 優：DB設計を支援する情報資源辞書システムの操作機能と実現法，情報処理学会論文誌，Vol. 33, No. 4, pp. 532-542 (1992).
17) 穂鷹良介：管理実体型の概念について，情報処理学会データベースシステム研究会，87-1, pp. 1-7 (1992).
18) 堀内 一：企業情報システムにおけるデータ中心手法導入の要件，情報処理学会論文誌，Vol. 33, No. 4, pp. 521-531 (1992).
19) 堀内 一：データ中心システム設計，オーム社 (1988).
20) 松尾比呂志：意味属性に基づくテキストベースの検索方式，情報処理学会論文誌，Vol. 32, No. 9, pp. 1172-1179 (1991).
21) 規格票の様式，日本規格協会，JIS Z 8301, pp. 18-27 (1990).
22) 情報資源スキーマ調査研究報告書，日本規格協会 (1990).
23) 先進ユーザがうかんだデータ項目名標準化の糸口，日経コンピュータ，9.14 号，pp. 105-113 (1987).

(平成4年 6 月17日受付)
(平成4年12月10日採録)

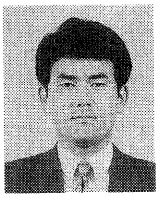
関根 純（正会員）

1958 年生．1980 年東京大学工学部計数工学科卒業．1982年同大学院修士課程修了．同年，日本電信電話公社横須賀通信研究所に入社．データベース管理システム，マルチメディアデータベースの研究実用化に従事．現在，データベース設計支援，情報資源管理の研究実用化を行う．情報規格調査会 SC 21/WG 3/RMDM+IRDS サブグループ委員．ACM 会員．



川下 満（正会員）

1951 年生．1975 年山口大学大学院工学研究科修士課程修了．同年日本電信電話公社入社．データベースの分散処理方式，マルチメディア情報検索等の研究実用化に従事．現在 NTT 情報通信網研究所にて，データベースの設計法，および，その支援システムの研究実用化を行っている．



町原 宏毅（正会員）

1963 年生．1987 年慶應義塾大学理工学部数理科学科卒業．同年，NTT 入社．主にデータベース管理システム，およびデータベース設計支援の研究実用化に従事．

中川 優（正会員）

昭和22年生．昭和45年大阪大学基礎工学科制御工学科卒業．昭和47年同大学院修士課程修了．同年，日本電信電話公社武蔵野通研入社．OS, DBMS の実用化，および，自然言語理解，知識処理の研究に従事．現在，NTT 情報通信網研究所データベース研究部にて，データベース設計法，情報資源管理の研究実用化に従事．主幹研究員．工学博士．人工知能学会，電子情報通信学会各会員．