# セルラーデータシステム（CDS）を用いた汎用データ管理システムの開発

児玉敏男†，國井利泰‡，関　洋一††

† 前田建設工業株式会社技術研究所（〒102-0071東京都千代田区飯田橋3-11-18光邦ビル４F）

‡ ㈱モルフォ（〒112-0004東京都文京区後楽2-6-1 飯田橋ファーストタワー31階）

†† ソフトウェアコンサルタント（〒191-0001東京都日野市栄町3-8-2）

E-mail: † kodama.ts@jcity.maeda.co.jp, ‡ kunii@ieee.org, kunii@acm.org, †† yseki@amber.plala.or.jp

建設分野への公共投資の減少、公的機関の入札における総合評価方式の導入等により、年々受注競争が激化する中、各建設企業は生き残りをかけ自社の強み弱みを客観的に分析した戦略的な経営判断が強く求められている。その際に、情報技術（IT）の戦略的利用は有効な手段であり、経営上重要度が高い多くのデータを柔軟に迅速に入出力・分析を行うことが肝要である。しかし、既存の一般的な方法で開発された業務アプリケーションでは、激しく変化する業務環境の中でユーザー要求の変化に対応できていない場合が多い。建設企業数社へのヒアリング調査結果より、経営上重要度の高い各業務におけるデータ管理上のユーザーの各要求は、１．異なるフォーマット（データ項目）の混在の許容、２．データ分析項目の柔軟な決定・追加・変更、３．語句の表記の揺れに対応したデータ検索、４．分析者の経験に依存しないデータ中心の分析、５．データファイルの一項目の値に複数の語句の入出力、であることが分かった。本研究では、柔軟なデータ処理を行うセルラーデータシステム（CDS）を用いて、上記１－５のユーザーの各要求を満足する汎用データ管理システムを開発した。

**キーワード：**セルラーデータシステム, 式表現, 汎用データ管理システム,データ分析

# The Development of a General-Purpose Data Management System Using Cellular Data System

Toshio Kodama†, Tosiyasu L. Kunii‡, Yoichi Seki††

† Maeda Corporation, Inc. (Koho Bldg. 4th floor, 3-11-18 Iidabashi, Chiyoda-ku, Tokyo, 102-0071 Japan)

‡ Morpho, Inc. (Iidabashi First Tower 31st floor, 2-6-1 Koraku, Bunkyo-ku, Tokyo, 112-0004 Japan)

†† Software Consultant (3-8-2 Hino-City, Tokyo, 191-0001 Japan)

E-mail: † kodama.ts@jcity.maeda.co.jp, ‡ kunii@ieee.org, kunii@acm.org, †† yseki@amber.plala.or.jp

Due to the reduction in public investment for construction projects and the introduction of a new bidding system for public entities, and while competition for orders is getting fiercer by the year, there is a great need for construction companies to make strategic business decisions based on an objective assessment of their company's strengths and weaknesses. To that end, the strategic use of information technology (IT) is an effective means, and it is essential that information important to the operation of the company be in/output and analyzed flexibly and promptly. However, business applications developed through the method currently in use are oftentimes unable to handle constantly changing user requirements. From consultations with a number of general contractors, we have learned that the data management requirements for the highest priority business tasks are:

1. intermixing of files with different attributes
2. flexible decisions, additions, and changes to data analysis items
3. data searches that accommodate variances in transcription
4. data-centered analysis that does not depend on the analyst's experience
5. input/output of multiple attributes for each data file

To satisfy the six user requirements above, our research has developed a general-purpose data management system employing the Cellular Data System, a flexible system of data processing.

***Key Words :*** *cellular data system, formula expression, a general purpose data management, data analysis*

## 1. はじめに

建設分野への公共投資の減少、公的機関の工事入札における総合評価方式の導入等により、年々受注競争が激化する中、各建設企業は生き残りをかけ自社の強み弱みを客観的に分析した戦略的な経営判断が強く求められている。同時に、業務をスリム化し必要コストを削減することも強く要求されている。その際に、情報技術（IT）

の戦略的利用は有効な手段となり得る。ITを利用したデータ管理の視点からその対策を俯瞰すると、入札結果データ、購買データ、顧客データ等の企業が扱う大量のデータを柔軟に迅速に入出力、分析を行うことがより重要であるが、しかし、既存の一般的なITを利用した業務の状況は、激しく変化する業務環境の中でユーザー要求の変化に対応できていない場合が多い。ITを利用するユーザーの立場からその理由を調査すると、多くのデータファイルからデータ出力を行うとき頻繁に発生する複雑な出力要求に柔軟に対応できないこと、同一単語の表記の相違（表記の揺れ）による出力データ不整合、データ入力設計の制限によるデータ入力漏れ、各業務アプリケーションや各組織で異なるデータフォーマットの相違によりデータ統合と総合的な分析が困難であること、データファイルの分類整理に柔軟性が欠如していること等が、多くのユーザーの業務生産性を大きく低下させていることが分かった。これらの課題を改善するには、データ管理の視点から以下のユーザーのシステム設計への各要求（要求１～要求５）を満たすことが条件であると考えられる。

要求１：業務AP開発時に、データファイルの項目を定義するのではなく、異なるフォーマットのファイルの混在を許容できるデータ管理を可能にすること

要求２：業務AP設計時に、ユーザーのデータ出力要求を定義するのではなく、データ出力時にユーザー自らがシステム利用時、柔軟に出力要求を決定できること

要求３：一つの単語に対して表現の統一を前提にせず、ユーザーが同義語を柔軟に定義しながらデータ出力時に同義語を考慮した出力を可能にすること

要求４：データ分析時に分析者の経験によらないデータ中心の分析が可能であること

要求５：データファイルの一つの項目の値として、複数の語句の入出力が可能であること

本論文では、上記の５つの要求を満足する汎用データ管理システムの開発を目的とする。

## 2. セルラーデータシステム（CDS）と既往の主な技術との比較

### ⑴ 開発体制

CDSは、セルラーモデル[1]を、式表現[2]と呼ばれる形式言語の一つを用いて設計・実装された新しいデータ処理システムである。CDSは、事象のデータモデリングで高い汎用性が発揮され、業務APシステム開発においてはユーザー要求を柔軟にデータ構造に射影することが可能になり、その結果、開発期間短縮・コスト削減に貢献する。

### ⑵ セルラーモデルと式表現

セルラーモデルは、以下の１～６の増加的モジュラー式抽象階層（Incrementally Modular Abstraction Hierarchy、以下IMAH）に沿って、サイバーワールドと実世界の構造をモデリングする[1]。

１．ホモトピーレベル
２．集合論レベル
３．トポロジー空間レベル
４．セル空間レベル
５．表現レベル
６．ビューレベル

サイバー空間上のサイバーワールドをモデリングするとき、IMAHに従って高いレベルからより規定されたレベルへ段階的に、サイバーワールドのプロパティーを定義する。

式表現は、形式言語の一つであり、オブジェクトとそれらの関係を、集合、順序集合、和、積というシンプルな形式で表現できる点が特徴である。式表現は、以下の文法から生成される[2]。

$G=(\{E,T,F,id\},\Sigma_L\cup\{\varepsilon,\varphi,+,\times,(,),\{,\}\},P,E)$

$P=\{E\to T|E+T+T\to F|T\times F+F\to(E)|\{E\}|id+id\to w\}$

()：集合、{}：順序集合、ε：単元、φ：零元、$\Sigma_L$：意味を表す語句の集合、w：語句（$\in\Sigma_L$）

このとき、Eを式、Tを項、Fを因子、idを識別子と呼ぶ。また、×や()、{}の括弧は矛盾が無い範囲で省略可能であるとする。

式表現rが表現する言語をL(r)とすると、L(r)は以下のように定義される。

1. $L(a)=\{a\}\ (a\in\Sigma_L)$
2. $L(\varepsilon)=\{\varepsilon\}$
3. $L(\varphi)=\{\}$
4. $L(r+s)=L(r)\cup L(s)$
5. $L(r\times s)=\{r\times s\}$

また、式表現 r, s, t, u は以下1-7の代数的構造を満たす。

1. $r+(s+t)=(r+s)+t,\ r\times(s\times t)=(r\times s)\times t$
2. $r+s=s+r$
3. $r\times\varepsilon=\varepsilon\times r=r$
4. $r\times\varphi=\varphi\times r=\varphi,\ r+\varphi=r$
5. $r\times(s+t)=r\times s+r\times t,\ (r+s)\times t=r\times t+s\times t$
6. $r+r=r$
7. $\{r+s\}\times\{t+u\}=r\times t+s\times u$

この式表現による、セルラーモデルの各レベルの空間の詳細設計については本稿では割愛し、参考文献[3]を参照されたい。

### ⑶ ユーザー要求に対するCDSと他の主要技術との比較

１章の各要求に応え得る汎用データ管理システムの開発には、既存の主な他の技術と比較して著者らによって

開発されたセルラーデータシステムの利用が有効であると考えられる。なぜなら、CDSはデータモデルのレベルで、１については空間の排他和の機能、２，３についてはトポロジー空間としての同義語空間の設計とその条件式検索機能への適用、４については集合処理の継承機能、５、６についてはトポロジー空間の継承機能、がサポートされているため、要求１～要求６を満たすデータモデルとしての汎用性の確保が可能であるからである。これに関して、既存の主なデータモデルとして、XML（eXtensible Markup Language）、OODB（Object Oriented DataBase）、RDB(Relational DataBase)を取り上げ、各技術とCDSが採用するデータモデルを表3.1に、前述の要求１～要求６への対応についてのデータモデリングにおける汎用性の比較を表3.2に示した[3]-[7]。

表2.1　主要な各技術が採用するデータモデル

| 技術 | データモデル |
|---|---|
| XML | タグによる木構造モデル |
| OODB | オブジェクト指向モデル |
| RDB | リレーショナルモデル |
| CDS | セルラーモデル |

表2.2　各要求への対応に対するCDSと主な技術との比較

| | XML | OODB | RDB | CDS |
|---|---|---|---|---|
| 要求１ | △ | × | × | ○ |
| 要求２ | △ | × | ○ | ○ |
| 要求３ | △ | △ | × | △ |
| 要求４ | × | × | × | △ |
| 要求５ | ○ | × | × | ○ |

表2.2において、各要求への各技術の対応の程度について、「○」は“データモデルで対応可能な機能を有する”ことを意味し、「△」は“データモデルではサポートしないが、容易な応用設計で対応可能である”ことを意味し、「×」は“データモデルではサポートされず、全てアプリケーションレベルでの機能開発が必要である”ことを意味する。

## 3. CDSを利用した汎用データ管理システムの開発

### (1)　システムの全体像と既存技術による開発方法との比較

本システムは、ユーザーインターフェース（WEBブラウザ）、AP ・WEBサーバー、アプリケーションプログラム、CDS、データベースから構成されるWEBアプリケーションシステムである。また、本システムにより、データベース設計、アプリケーション開発を行う必要なく、業務データをそのまま式データに変換され取り込まれ、柔軟なデータ検索が可能になる。ユーザーのデータ入力時、表構造データ、同義語・関連語データ、階層構造データ等を各コンバートプログラムにより、式データに自動変換し、式表現の統合機能により式データが統合され、データストレージ（RDB）に保存される。データ出力時は、CDSの条件式検索機能に同義語・関連語処理が組み込まれていて、同義語・関連語を考慮した柔軟なデータ検索が可能である。また、表構造データについては、ブラウザへの出力とともにCSVファイルとしての出力可能にすることで、出力後のデータ加工を容易にする（図4.1）。本システムの機能は、表構造データ分析機能、階層構造データ管理機能に大別される。表構造データ分析機能は、フォーマットの異なる表構造のデータを統合し、汎用的なデータ検索を可能にする機能である。階層構造データ管理機能は、複雑な階層構造データの柔軟な管理を可能にする機能である。

本システムは、クライアント端末からネットワーク環境とブラウザさえあればどこからでも要求・応答を行うことのできるWEBアプリケーションシステムとしての設計にした。また、開発言語は色々なOSに対応可能なJAVAを、堅牢性の確保のためにリレーショナルデータベース（RDB）をデータストレージとして採用した。

### (2)　ユーザーの各要求とシステム機能との対応

2章(4)で述べたデータ管理システムへの各要求と本システムの機能設計との対応を次に述べる。

要求１：項目が異なる表構造データの各ファイルが混在できること

対応機能：表構造データの式データへの変換（(4)のa)）、和演算による式データの統合機能（(4)のc)）、セル空間分割関数による整形出力機能(4)のg)）で対応

説明：表構造データはセル空間の式データとして表現され、複数のセル空間の式データは和演算(+)により容易に統合可能である。

要求２：データ出力時にユーザー自らがシステム利用時に出力要求を決定できること

要求３：ユーザーが同義語・関連語を柔軟に定義でき、それらを考慮した出力を可能にすること

対応機能：同義語・関連語処理関数による条件式検索機能への同義語・関連語の適用（(4)のf)）で対応

説明：同義語・関連語データの式データによる設計（(4)のb)）、表構造データに対する条件式検索((4)のe)）、同義語・関連語処理関数による条件式検索への同義同義語・関連語は、集合またトポロジー空間の式データとして表現でき、また同義語・関連語は集合論の論理和の演算として解釈できるので、条件式検索に適用可能である。

要求４：データ中心のデータ分析が可能であること

対応機能：識別子数取得関数による自動集計機能（(4)のh）で対応

説明：自動集計機能により分析項目となる語句の候補が

自動的に出力され、各語句の頻出順にグラフが作成される。

要求５：データファイルの一つの項目の値として、複数の語句を入出力が可能であること

対応機能：表構造データの式データへの変換（(4)のa)）で対応

説明：表構造データを式データに変換時、各項目の値に複数の値を入力することができる。

### ⑶　表構造データ分析機能の設計

a)　表構造データの式データへの変換

分析対象である表構造データの入力時のファイル形式は、ユーザーフレンドリーなファイルフォーマットのCSVファイルとする。CSVファイルには、１行目に各項目を、２行目以降に各レコードを入力する。このとき、一カ所に複数の値や値間の関連が入力可能である。

作成したCSVファイルをサーバーにアップロードし、属性と属性値の始めの行を設定（図3-2）した後、コンバータプログラムにより、次の式3-1の属性を持つセル空間としての式データが作成される。

$$filename\{\sum 属性_i\}(\sum(\sum\{属性値_i\})) \quad \text{（式 3-1）}$$

例を示す。

file1{A+B+C+D+E+F+G}(1{(a11+a12)+b1+c11×c12+d1+e1+f1+g1}+2{a2+b2+…+g2}+3{a3+b3+…+g3}+4{a4+b4+…+g4}+…+10{a10+b10+….+g10})

（式 3-2）

b)　同義語・関連語データの式データによる設計

同義語・関連語を考慮した検索を行う要求があるとき、各ユーザーでそれぞれ同義語・関連語データを登録する必要がある。（関連語とは、語句の集合とそれらを総称する語の組み合わせとする。例えば、ユーザーによって、語の集合「鹿島建設」、「大成建設」、「清水建設」と総称する語「ゼネコン」は、関連語として定義される。）同義語・関連語の登録は、登録画面において、ユーザーが語句データをカンマ(,)で区切り入力し、入力ボタンを押下して同義語・関連語のグループを登録していく。登録されたデータはサーバーに送信された後、コンバータプログラムにより以下の式3-3の集合またはトポロジー空間の式データが作成される。

SYNONYM($\sum_i$($\sum$同義語の要素$_i$の語))+RELATION($\sum_j$ 総称を表す語($\sum$関連語の要素$_j$の語))

（式 3-3）

例を示す（式3-4）。

SYNONYM((トヨタ自動車+TOYOTA)+(NISSAN+日産自動車)+(HONDA+ホンダ+本田技研)+…)+RELATION(産業(自動車+機械+繊維+建設)+…)

（式 3-4）

c)　和演算による式データの統合機能

各ファイルから変換された複数の式データをCDSのAPIを使用し、和演算により統合を行う。これにより、フォーマットの異なる各ファイルのデータが一つの同じデータストレージへの保管が可能になる。1~nまでのファイルが統合された式データを$fml_n$、n+1番目に統合するファイルの式データを$tm_{n+1}$、式データを統合する関数を$f$とすると$fml_{n+1}$は以下（式3-5）になる。

$$fml_{n+1} = f(fml_n, tm_{n+1}) = fml_n + tm_{n+1} \quad \text{（式 3-5）}$$

d)　RDBへのマッピングのための絶対位置情報変換関数

前述（2.1.2）の通り、式は項の和からなり項は因子の積からなる。式におけるm番目の項のn番目の因子に対し(m, n)を因子の位置情報と呼ぶことにする。ここで、CDSのAPIを使用し、式データに対して以下（式3-6）の写像により像を求める。

$$f : fml \rightarrow \sum_{k=1}^{n} \sum_{k=1}^{m} (m,n) \times fct$$

$fct$が()括弧形の場合は再帰的に繰り返し写像を行う

（式 3-6）

このとき、関数$f$を絶対位置情報変換関数と呼ぶ。上記（式3.2.5）の写像によって、式データにおける全ての因子を位置情報と値の組に変換できるので、RDBのテーブルへのマッピングが可能になる。

e)　表構造データに対する条件式検索

条件式検索は、CDSのAPIを利用して、入力された条件式に合致する式データを出力する。検索対象のセル空間が統合された式データを$\sum filename_i\{\sum 属性_j\}(\sum(\sum\{属性値_k\}))$、ユーザーの条件式を$p+q$、CDSの条件式検索関数を$f$とすると、出力される式データは以下（式3-7）である。

$$f(\sum filename_i\{\sum 属性_j\}(\sum(\sum\{属性値_k\})), p+q) = \sum filename_i\{\sum 属性_j\}(\sum(p\text{ を含むレコード})) + \sum filename_i\{\sum 属性_j\}(\sum(q\text{ を含むレコード})) \quad \text{（式 3-7）}$$

f) 同義語・関連語処理関数による条件式検索機能への同義語・関連語の適用

条件式検索において、Aを含む式データを取得するときAとaの意味が同値であるならば、条件式は(A+a)と表すことができる。CDSのAPIを利用し、入力する条件式の各識別子に対し、定義した同義語・関連語データを参照して、条件式に同義語・関連語を組み込む。同義語・関連語の式データが以下（式3-8）のように定義されていたとする。

SYSNONYM((平成１９年+2007)+(平成２０年+2008)+(平成２１年+2009)+(事件+事故+アクシデント)…)+RELATION(春(3月+4月+5月)+夏(6月+7月+8月)+秋(9月+10月+11月)+冬(12月+1月+2月))

（式 3-8）

このとき、入力する条件式を例えば、平成２０年×夏×事件（“平成２０年”かつ“夏”かつ“事件”の意味）とすると、同義語・関連語処理関数の適用により、入力する条件式は内部で以下（式3-9）のように変換される。

変換前：平成２０年×事件
変換後：(平成２０年+2008)×(夏+6月+7月+8月)×(事件+事故+アクシデント)

（式 3-9）

g) セル空間分割関数による整形出力機能

条件式検索で得られた式データを出力する時、CDSのAPIを使用し、各ファイルで選択した属性の出力を可能にする。これを整形出力と呼ぶことにする。

ユーザーが整形出力を行うとき、入力したセル空間の式データが分割され、指定した属性の列を現すセル空間の式データが取得される。入力されたセル空間の式データを*filename*{∑*属性*$_i$}(∑(∑{*属性値*$_i$}))、指定する属性を*属性*$_n$,*属性*$_m$,…、CDSのセル空間分割関数を*f*とすると、取得されるセル空間の式データは以下（式3-10）である。

$$f(filename\{\sum 属性_i\}(\sum(\sum\{属性値_i\})),(属性_n+属性_m+\ldots)) = filename\{属性_n+属性_m+\ldots\}(\sum(属性値_n+属性値_m+\ldots))$$

（式 3-10）

例を示す（式3-11）。

f(file1{A+B+C+D+E+F+G}(1{(a11+a12)+b1+c11×c12+d1+e1+f1+g1}+2{a2+b2+…+g2}+3{a3+b3+…+g3}+4{a4+b4+…+g4}+…),(A+B))
=file1{A+B}(1{(a11+a12)+b1}+2{a2+b2}+3{a3+b3}+4{a4+b4}+…)

（式 3-11）

h) 識別子数取得関数による自動集計機能

条件式検索で得られた式データから、CDSのAPIを使用し、各語句の出現頻度の検出を可能にする。これを自動集計機能と呼ぶことにする。また、検出頻度から自動的に簡単なグラフの作成も可能にする。

ユーザーが自動集計を行うとき、入力したセル空間の式データが表現レベルの数値子式の式データに変換された結果が取得される。入力されたセル空間の式データを*filename*{∑*語*}(∑(∑{*語*}))、CDSの識別子数取得関数を*f*とすると、取得される式データは以下（式3-12）である。

$$f(filename\{\sum 識別子_i\}(\sum(\sum\{識別子_j\}))) = 識別子1\times l+識別子2\times m+識別子3\times n+\ldots$$

(l,m,n…は数を表す識別子)

（式 3-12）

i) 検索結果のMSエクセルへの出力機能

条件式検索による検索結果を、ユーザーフレンドリーなMSエクセルでの出力を可能にする。MSエクセルに出力することでユーザーによるデータの2次加工が容易になる

## 4. システムの検証

### (1) ユーザーの各要求と業務適用との対応

検証を行う業務として、本章(3)で総合評価方式の入札結果データ分析の課題を、本章(4)で資材調達実績データ分析の課題をそれぞれ取り上げる。また、本システムを適用する各業務と、システムへのユーザーの各要求の対応を以下（表4.1）にまとめた。

表4.1 システムを適用する業務とユーザーの各要求との対応

| 適用する業務 | ユーザー要求の NO.（1章） |
|---|---|
| 総合評価方式の入札結果データへの適用 | １，２，３，４，５ |
| 資材調達実績データ分析の課題 | ２，３，５ |

### (2) テスト環境

本システムは、クライアント端末からネットワーク環境とブラウザさえあればどこからでもサーバーに要求・応答を行うことのできる WEB アプリケーションシステムである。システムの処理速度は、サーバーマシーンの機能に依存するが、マシーンのスペック概要は以下である。

CPU： 2.83GHz（デュアルコア）
メモリ：4GB

OS：Red Hat Enterprise Linux 5

また、本システムのソフトウェア構成を以下に示した。

WEB サーバー：apache2.2.2

AP サーバー：Tomcat5.5.4

JAVA：JDK1.5.015

RDB：mysql5.1

### (3) 総合評価方式の入札結果データへの適用

a) ファイルの取得

先ず、国土交通省各９整備局、NEXCO３社（中日本、東日本、西日本）の各総合評価方式の入札結果データファイル（2005年4月～2008年9月までの217のファイル、200万レコード、600MB）が各WEBサイトからダウンロードされ、CSVファイルに変換、名前を付けて保存される。前述ようにデータファイルの各項目はそれぞれ異なっている

b) サーバーへのアップロード

各CSVファイルをサーバーにアップロードする。

１．ファイルアップロード画面でCSVファイルを選択・サーバーにアップロードすると、CSV解析画面に遷移

２．CSV解析画面で、項目・値の行を選択し解析を行う。

３．１，２を繰り返す。

c) 同義語の設定

入力した総合評価方式の結果データにおいて、一般的に同義語と考えられる以下（表4.2）の語句を入力画面から登録する。

表4.2 入札結果データの同義語の例

| NO | 同義語 |
|---|---|
| 1 | 決定，落札 |
| 2 | H18，2006 |
| 3 | H19，2007 |
| 4 | H20，2008 |
| 5 | （株），株式会社 |
| 6 | JV，共同企業体 |

d) 条件式検索を利用したデータ分析

条件式検索を行い、必要なデータを出力しブラウザ上に表示する。必要な場合は出力データをCSVファイルとしてダウンロードし、2次加工する。次の各出力要求（出力要求１～出力要求４）に対して、本システムにおける条件式検索を利用したデータ分析例を上げる。

出力要求１：

「全国の総合評価方式のWTO対象工事案件において、入札件数、落札件数、受注率を大手４社の平均と前田建設を比較したい」というデータ分析要求

検索１－１：

まず、大手４社のWTO対象工事への入札件数を調べるため、検索窓に条件式を"(鹿島建設 OR大成建設 OR 清水建設 OR 大林組)WTO"と入力し、検索すると、件数は599件（処理時間0.63秒）であった。よって、大手４社のWTO対象工事への入札件数の１社平均は149件（599/4）であることが分かった。

検索１－２：

次に、大手４社のWTO対象工事への落札件数を調べるため、検索窓に条件式を"(鹿島建設 OR大成建設 OR 清水建設 OR 大林組) WTO 落札"と入力し、検索を行うと、件数は96件（処理時間0.22秒）であるという結果が得られた。

検索１－３：

次に、前田建設のWTO対象工事への入札件数を調べるため、検索窓に条件式を"前田建設工業 WTO"と入力し、検索を行うと入札件数は164件（処理時間0.32秒）という結果が得られた。

検索１－４：

次に、前田建設のWTO対象工事への落札件数を調べるため、検索窓に条件式を"前田建設工業 WTO 落札"と入力し、検索を行うと、入札件数は22件（処理時間0.20秒）であるという結果が得られた。

よって、前田建設のWTO対象工事の受注率（落札件数/入札件数）は、13.4%（22/164）であることが分かった。

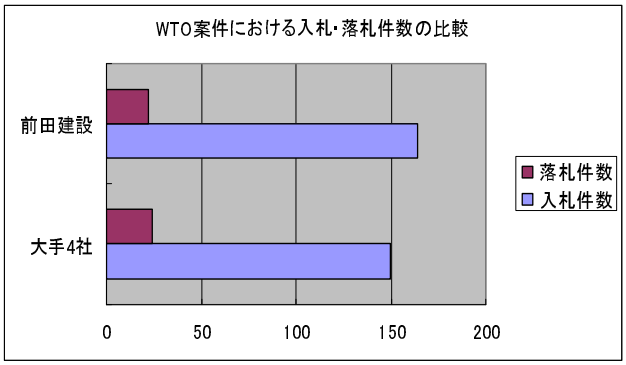


図 4.1 WTO 案件における入札・落札のグラフ

出力要求２：

「北海道の工事案件における入札件数、落札件数、受注率を大手４社と前田建設を比較したい」というデータ分析要求

入力２－１：

北海道地方整備局は他の地方整備局とは異なり、次の各地域で各建設部に管轄が分かれている。よって、各レコードに「北海道」という語句が入っていないので、関連語登録画面で関連語として表2.3の語句を登録する必要がある。

表 4.3 入札結果データの関連語の例

| 総称語 | 関連語 |
|---|---|

| | |
|---|---|
| 北海道 | 札幌，函館，小樽，旭川，室蘭，釧路，帯広，網走，留萌，稚内，石狩川 |

検索２－１：
先ず、北海道の工事案件における大手４社の入札件数を調べるために、検索窓に条件式を“(鹿島建設 OR 大成建設 OR 清水建設 OR 大林組) 北海道”と入力し、検索を行うと、入札件数121件という結果が得られた。同様に、落札件数を調べるために、前式に”落札”を加えて検索すると24件（処理時間0.15秒）であった。よって、大手４社の北海道の工事案件において受注率(落札件数/入札件数=24/121)は、19.8%であることが分かった。

検索２－２：
北海道の工事案件における前田建設の入札件数を調べるために、検索窓に条件式を“前田建設 北海道”と入力し、検索を行うと、入札件数は97件（処理時間0.30秒）であった。次に、その中で、落札件数を調べるために、前式に識別子”落札”を加えて検索すると、そのうち落札件数は9件であった。よって、前田建設の北海道の工事案件において、受注率（落札件数/入札件数）は9.3%（9/97）であることが分かった。

これらから、北海道地方の工事案件の入札・落札件数について、前田建設と大手４社と比較した結果をグラフは図2.2である。

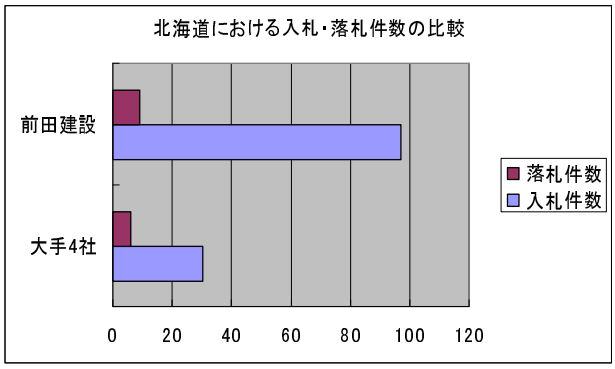


図 4.2 北海道における入札・落札件数の比較

出力要求３：
「東北地方のＷＴＯ対象工事案件において、入札件数や落札件数が多い企業を比較したい。またそれらの受注率を知りたい」というデータ分析要求

検索３－１：
東北地方のＷＴＯ対象案件で入札件数が多い企業名を取得するため、検索窓に条件式を“東北 WTO”と入力し検索を行い、自動集計機能を使用し出現頻度を整理し、上位20件（処理時間1.31秒）のみ抽出すると以下（表2.4）になった。

検索３－２：
次に、東北地方のＷＴＯ対象案件で落札件数が多い企業名を取得するため、検索窓に条件式を“東北 WTO 落札”と入力し、検索を行い、自動集計機能を使用し、出現頻度を整理し、上位から抽出する（処理時間0.91秒）と以下（表4.5）になった。

表 4.4 東北地方の WTO 対象案件の入札件数
（上位 20 位）

| 企業名 | 入札件数 |
|---|---|
| 三井住友建設㈱ | 17件 |
| 飛島建設㈱ | 16件 |
| 前田建設工業㈱ | 15件 |
| ㈱大林組 | 14件 |
| 青木あすなろ建設㈱ | 14件 |
| 大成建設㈱ | 14件 |
| ㈱奥村組 | 13件 |
| ㈱銭高組 | 13件 |
| ㈱間組 | 12件 |
| 鹿島建設㈱ | 12件 |
| 清水建設㈱ | 12件 |
| 五洋建設㈱ | 12件 |
| ㈱鴻池組 | 12件 |
| 日本国土開発㈱ | 11件 |
| 佐藤工業㈱ | 11件 |
| ㈱竹中土木 | 10件 |
| 西松建設㈱ | 10件 |
| ㈱不動テトラ | 10件 |
| ㈱熊谷組 | 9件 |
| 戸田建設㈱ | 9件 |

表4.5 東北地方のWTO対象案件の落札件数
（上位から）

| 企業名 | 落札件数 |
|---|---|
| 三井住友建設㈱ | 5件 |
| ㈱奥村組 | 3件 |
| ㈱間組 | 3件 |
| 飛島建設㈱ | 3件 |
| 青木あすなろ建設㈱ | 2件 |
| 大成建設㈱ | 2件 |
| ㈱フジタ | 2件 |
| ㈱JFEエンジニアリング | 2件 |
| 豊国工業㈱ | 2件 |
| （その他 16 社） | 1件 |

東北地方のWTO対象案件の入札件数上位20社を受注率

の順で整理すると以下（表4.6）になる。

表 4.6　東北地方の WTO 対象案件における建設企業各社の受注状況（受注率上位 20 社）

| 企業名 | 入札件数 | 落札件数 | 受注率 |
|---|---|---|---|
| 三井住友建設㈱ | 17 件 | 5 件 | 29.4% |
| ㈱間組 | 12 件 | 3 件 | 25.0% |
| ㈱奥村組 | 13 件 | 3 件 | 23.1% |
| 飛島建設㈱ | 16 件 | 3 件 | 18.8% |
| 青木あすなろ建設㈱ | 14 件 | 2 件 | 14.2% |
| 大成建設㈱ | 14 件 | 2 件 | 14.2% |
| ㈱竹中土木 | 10 件 | 1 件 | 10.0% |
| 鹿島建設㈱ | 12 件 | 1 件 | 8.3% |
| 清水建設㈱ | 12 件 | 1 件 | 8.3% |
| ㈱銭高組 | 13 件 | 1 件 | 7.7% |
| 前田建設工業㈱ | 15 件 | 1 件 | 6.7% |
| ㈱大林組 | 14 件 | 0 件 | 0.0% |
| 五洋建設㈱ | 12 件 | 0 件 | 0.0% |
| ㈱鴻池組 | 12 件 | 0 件 | 0.0% |
| 日本国土開発㈱ | 11 件 | 0 件 | 0.0% |
| 佐藤工業㈱ | 11 件 | 0 件 | 0.0% |
| 西松建設㈱ | 10 件 | 0 件 | 0.0% |
| ㈱不動テトラ | 10 件 | 0 件 | 0.0% |
| ㈱熊谷組 | 9 件 | 0 件 | 0.0% |
| 戸田建設㈱ | 9 件 | 0 件 | 0.0% |

e)　考察

業務APシステム開発において、従来のシステム開発の手法では、１．要求分析、２．分析に基づくデータベース設計、２．各機関の入札結果データの各ファイルからのデータベースへのコンバートプログラムの開発、４．入出力設計、５．入出力プログラム実装・テスト、６．システムテスト、を行う。ここで、各ファイルでフォーマットが少しずつ異なるので全ファイルの全項目のデータを活用できるわけではない。さらに、その後、他の機関からの入札結果データを合わせて分析を行うという要求が生じたときには始めから全開発プロセスを見直す必要がある。これに対して、本システムを利用すれば、操作画面からユーザーが、１．各機関のCSV形式のファイルのアップロード、２．項目・値行の設定、３．同義語・関連語の設定、４．出力のための整形設定、と順に行えばよい。他の機関のフォーマットが異なるファイルを利用したいときも１～４の操作を行えばよい。また、このとき、決められたデータベース設計が無いので全ファイルの全データが活用できる。さらに、データ出力の仕様もも条件式検索として検索機能が汎用化されているので、仕様変更に伴うメンテナンスも最小限に抑えることができる。

出力要求１では、各同義語の設定した条件式検索により、柔軟に検索の条件を発展させることで、全国のWTO対象工事における前田建設と大手建設4社との入札・落札件数、受注率の比較について容易に出力することができた。出力要求２では、出力要求に合わせて、北海道に関連する各地名を関連語として登録することで、北海道の入札結果データを柔軟に出力でき、前田建設と他社との比較が可能になった。このように、ユーザーは合目的的ではなく試行錯誤を繰り返しながらデータ出力・分析が可能なので、企業の受注戦略のためのデータ分析に有益であると考えられる。また、一般的にデータ分析においては、分析者の先入観や経験が公平な分析を妨げる場合があるが、出力要求３で使用した自動集計機能により、入札結果データにある全語句の各出現頻度が自動的に計算されるので、データ中心の分析を行うことが可能になった。表2.6の分析結果は、従来の方法ではなされなかった結果であり、より客観的なデータ分析であることが分かる。

また、通常の検索に要した処理時間と出力件数の関係を図4.4に示した。これらの前述のような通常の分析処理では要した処理時間はほぼ1秒以下であり、迅速な業務処理に十分に対応可能と考えられる。

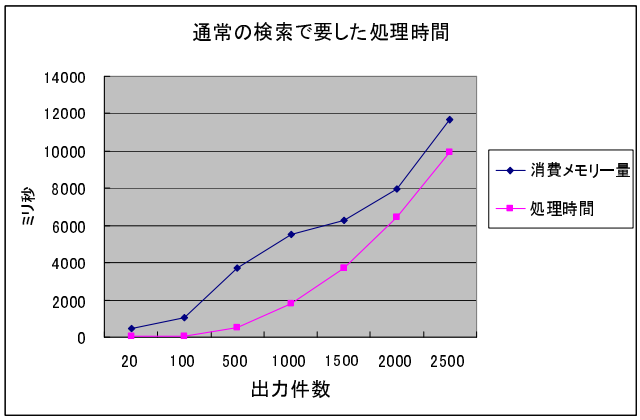


図 4.4　検索に要した処理時間と出力件数の関係

(4)　資材調達実績データ分析への適用

a)　データ入力と同義語登録

本節では、前田建設調達部が扱っている資材調達実績データ（H16～H19）の中のガラスデータの分析を行う。先ず、ガラスデータのファイル（約5,000レコード、1.2MB）を操作画面からサーバーへアップロードし、入力する。入力したガラスデータにおいて、同義語と考えられる語句を入力画面から登録する必要があるが、表記の揺れの幅が非常に大きくてユーザーが同義語全てを事前に把握できない場合は、条件式検索の差分を取得する機能（-）を利用し、同義語を設定しながらファイル中の全データから同義語と思われるデータを削除していく。これを、「表記の揺れの吸収」と呼ぶことにする。以下

より、ガラスデータの表記の揺れの吸収する例を上げる。

先ず、操作画面の検索窓に、ファイル名である”ガラスデータファイル”を入力し、ガラスデータのファイル中の全レコードを出力する。

次に、出力結果をから、ガラスの名称に関する表記の揺れを確認するため、“ガラスデータ　－フロートガラス”で検索を行い、全レコードから“フロートガラス”と書かれているレコードを削除する。

同様に、出力結果から “フロート板ガラス”は“フロートガラス”と同義語であると予想されるので、“フロートガラス”の同義語として設定する。

そして、“ガラスデータ　－フロートガラス”で再検索を行い、全レコードから“フロートガラス”、“フロート板ガラス”と書かれているレコードを削除し、出力結果を確認する。出力結果から、“フロートガラス”（‘ー’が‘ー’になっている）と表記されていることが分かったので、これもの同義語として設定する。

これらを繰り返すことで、“フロートガラス”の同義語は以下（表4.7）であることが分かった。このようにして、値の表記の揺れを吸収することができる。同様にして、ガラスの“厚さ”の値、“使用㎡制限”の値に関する表記も次のような同義語（表2.8、表2.9）があることが分かった。

表4.7　ガラスデータにおける”フロートガラス”の同義語の例

| 同義語 |
|---|
| フロートガラス，フロート板ガラス，透明フロートガラス，フロートガラス，フロートカ゛ラス |

表4.8　ガラスデータにおける「厚さ」値の同義語の例

| NO | 同義語 |
|---|---|
| 1 | t3，厚3.0，厚3，T３，t=3.0，t=3 |
| 2 | t4，厚4，t=4.0，T-4 |
| 3 | t5，厚5，厚さ5，厚５，T５，t=5.0，t=5，t-5 |
| 4 | t6，厚６，厚6，T６，t=6.0，t＝６．０，t=6 |
| 5 | 厚８，t８，厚8，t=8，T-8，T8，t8 |
| 6 | t12，厚12，厚１２，T１２，t=12 |
| 7 | t15，厚15，t１５ |
| 8 | t19，厚19，厚１９ |
| 9 | T6.8，6.8mm |

表4.9　ガラスデータにおける「使用㎡制限」値の同義語の例

| NO | 同義語 |
|---|---|
| 1 | 2.18m２，２．１８m２，2.18m2，２．１８㎡，2.18㎡ |
| 2 | ４．４５m２，4.45m2，４．４５㎡，4.45㎡，4.45m |
| 3 | ６．８１m２，6.81㎡，６．８１㎡，6.81m2 |
| 4 | ９．０９㎡，9.09㎡ |

b)　条件式検索を利用したデータ分析

前田建設の業務上、実際にあった以下の出力要求（出力要求１，出力要求２）に対し、本システムの条件式検索機能を利用したデータ分析の例を上げる。

出力要求１：

「各支店別（北海道支店、東北支店、関東支店、北陸支店、中部支店、関西支店、中国支店、九州支店の8支店）、また全社でフロートガラスの平均単価を知りたい」というデータ分析要求がある場合

検索１－１：

各支店のフロートガラスの平均購買単価を取得するため、検索窓に各条件式　”ガラスデータ　フロートガラス　北海道支店”、”ガラスデータ　フロートガラス　東北支店” 、を入力し検索を行い、その結果をエクセルに出力し、項目「単価」の平均値を求める。

検索１－２：

同様に、全社の平均購買単価を取得するため、条件式　”ガラス関係データ　フロートガラス"を入力し、検索を行い、平均値を求める。得られたデータをまとめると下の表2.10のようになった。

表4.10　各支店のフロートガラスの平均購買単価

| 支店 | 単価 |
|---|---|
| 北海道支店 | 1700円 |
| 東北支店 | 964.5円 |
| 北陸支店 | （該当レコードなし） |
| 東京支店 | 2006.71円 |
| 中部支店 | 1546.67円 |
| 関西支店 | 1528.12円 |
| 中国支店 | 1810.46円 |
| 四国支店 | （該当レコードなし） |
| 九州支店 | 1522.94円 |
| 全支店 | 1914円 |

出力要求２：

「フロートガラスの厚さ毎（3,4,5,6,8,12,15,19ミリ）に全社の平均単価を知りたい」というデータ分析要求がある場合

検索２－１：

フロートガラスの厚さ毎の平均購買単価を取得するため、検索窓に各条件式　"ガラス関係データ　フロートガラス　t3"、”ガラス関係データ　フロートガラス　t5”、…を入力し検索を行い、結果をエクセルに出力し、項目“単価”の平均値を求める。

表2.11　フロートガラスの厚さ毎の平均購買単価

| 厚さ | 単価 |
|---|---|
| 厚３ミリ | 664.6円 |
| 厚４ミリ | 951.7円 |
| 厚５ミリ | 1069.6円 |
| 厚６ミリ | 1521.8円 |
| 厚８ミリ | 2134.0円 |
| 厚１２ミリ | 3193.3円 |
| 厚１５ミリ | 5033.3円 |
| 厚１９ミリ | 6860.0円 |

c)　考察

・資材調達データ分析の現場では、データファイルにおける同一単語に対する表記の揺れ、曖昧な項目定義が、分析業務上の非常に大きな問題になっている。しかし、本システムを利用すれば、同義語を適用した条件式検索によりデータの差分を求めることが可能であり、また一項目の値として複数の語句が入出力可能なので、同義語を推測して表記の揺れを吸収しながら作業を進めることができる。
・ガラスデータにおいて、表記の揺れを吸収した後に条件式検索を行い、要求１では支店毎に、要求２では厚さ毎に容易に平均購買単価を算出することができ、既存の手法では現実的に難しい資材調達実績データの平均購買単価の算出という課題を解決することができる。
・本システムでは、検索結果をユーザーフレンドリーなCSVファイルにダウンロードできるので、その後のデータ加工が非常に容易になる。
・分析対象ではないデータが多く含まれている場合（上記の例ではガラスデータ以外の資材調達実績データ）の分析業務においては、それらの分析対象外の各レコードIDを排除すべき同値類のレコードとして同義語設定することで、差分を取得する条件式検索により検索結果から対象外データを排除していくことで対応可能である。
・本システムでのデータ検索に要した各処理時間は全て0.5秒以下であり、十分実際の業務に対応可能である。

## 5.　結論

ITを利用したデータ管理の視点からその対策を俯瞰すると、入札結果データ、購買データ、顧客データ等の企業が扱う大量のデータを柔軟に迅速に入出力、分析を行うことがより重要であるが、しかし、既存の一般的なITを利用した業務の状況は、激しく変化する業務環境の中でユーザー要求の変化に対応できていない場合が多い。ITを利用するユーザーの立場からその理由を調査すると、多くのデータファイルからデータ出力を行うとき頻繁に発生する複雑な出力要求に柔軟に対応できないこと、同一単語の表記の相違（表記の揺れ）による出力データ不整合、データ入力設計の制限によるデータ入力漏れ、各業務アプリケーションや各組織で異なるデータフォーマットの相違によりデータ統合と総合的な分析が困難であること、データファイルの分類整理に柔軟性が欠如していること等が、多くのユーザーの業務生産性を大きく低下させていた。本研究では、これらの問題を改善するため、著者らによって開発された新しいデータ処理システムであるセルラーデータシステム（CDS）を用いて汎用データ管理システムが開発された。そのシステムを前田建設工業における実業務に適用することで、前述のデータ管理上の６つのユーザーの各要求に対し、その有効性が確認された。

参考文献

[1] T. L. Kunii and H. S. Kunii, “A Cellular Model for Information Systems on the Web -Integrating Local and Global Information-”, Proceedings of 1999 International Symposium on Database Applications in Non-Traditional Environments (DANTE’99), November 28-30, 1999, Heian Shrine, Kyoto, Japan, Organized by Research Project on Advanced Databases, in cooperation with Information Proceeding Society of Japan, ACM Japan, ACM SIGMOD Japan, pp. 19-24, IEEE Computer Society Press, Los Alamitos, California, U.S.A.

[2] Toshio Kodama, Tosiyasu L. Kunii, Yoichi Seki, “A New Method for Developing Business Applications: The Cellular Data System”, *In Proc of CW'06*, pp. 65-74, IEEE Computer Society Press.

[3] Bernadette Farias Lósio, Ana Carolina Salgado, Luciano do Rêgo GalvÐo, “Conceptual modeling of XML schemas”, *In Proc. of WIDM'03*, ACM Press, pp.102-105, 2003.

[4] Giovanna Guerrini, Marco Mesiti, Daniele Rossi, “Impact of XML schema evolution on valid documents”, *In Proc. of WIDM'05*, ACM Press, pp.39-44, 2005.

[5] Setrag Khoshafian, “Object-Oriented Databases” pp. 132-142, John Wiley & Son, 1993.

[6] Takashi Washio, Hiroshi Motoda, “State of the art of graph-based data mining”, *In ACM SIGKDD Explorations Newsletter*, ACM Press, pp.59-68, 2003.

[7] Toshio Kodama, Tosiyasu L. Kunii and Yoichi Seki, "A Condition Formula Search", Proceedings of 2008 SIWN Congress(22-24 July 2008, Glasgow, UK), and also in International Journal of Communications of The Systemics and Informatics World Network (SIWN), Co-Sponsored by IEEE Systems, Man, and Cybernetics Society, pp.39-44, The Systemics and Informatics World Network (SIWN), UK, 2008.