KG-EDENS;
関西学院大学　経済データ抽出システム[1]の構築[2]

豊　原　法　彦

Our purpose of this paper is to introduce the data extraction system of the Japanese macro economic data(kg-edens). We designed this system for novice and veteran users, so we can use this system through homepage style interface without a code table and traditional comannd line. Furthermore, by using a free language under GPL, a central part has a wide portability from unix-base workstations to personal computers that run gawk.

JEL : C82; C87
Key Words : Internet-based methods

1. はじめに

経済学部の教育・研究では、まず現実を数量的に把握しそこから抽象化することにより、理論モデルを構築した後でさらにその妥当性について実証的なデータを用いて検討するという手法の重要性は、情報教育の進展とあいまって、一層高まっている。日経データ社より発売されている経済マクロデータからユーザが指定する項

1) Kwansei Gakuin Economic Data Extraction for Network System

2) 本システムは、その開発段階において旧 情報処理研究センター（現 情報メディア教育センター）、産業研究所の技術的・資金的サポートを受けている。

目、期間のデータを抽出するためのシステムは、関学では汎用ホスト[3]上で以前より開発されていた。すべての処理をサーバ上で行うというトータルシステムと抽出のみサーバに依存しそれ以降を端末で行うという 2 通りの方法が存在したが、端末台数不足、厳密な端末管理に加えて、前者の場合には操作の特殊性、後者の場合には抽出データの転送手続きの繁雑さ、さらには端末サイドのアプリケーションソフトの能力などの問題が指摘された。

いわゆる 1980 年代後半からのダウンサイジングという潮流に従って、学内にもワークステーションが多数導入され、その中に統計関係のサーバも含まれており、さらに全学的にネットワークが整備されたことから、日経 NEEDS のマクロデータのネットワークに対応した抽出システムを開発することとなった[4][5]。その際に考慮した点は以下の 3 つである。1) ユーザが利用する際に、サーバであるワークステーションとクライアントであるパソコンの間をできる限りシームレスにすること。2) 開発費用を低廉に押さえ、ライセンス料など管理が繁雑にならないようにした上で、メンテナンスなどを容易にすること、3) ライセンスなどの問題が許されればパソコンなどで運用できるようにすること。

データベースなどの特定のアプリケーションに依存した環境では、操作性やアクセススピードなどのメリットを認めつつも、そのソフトの制限に拘束される可能性があることやそのためにサーバ、クライアントとも追加的な投資が必要であることを考慮して、本システムは繰り返しデータ抽出を行うベテラン層から、初歩的な実証分析を始めたノービス層まで簡単なオペレーションで利用できることを目標に開発された。まず、ベテランユーザにはコマンドラインからの利用を想定し、繰り返し抽出を容易にするためにリダイレクトによる入力も考慮している。また、ワークステーションでもパソコンでも最少の変更で利用できるようにするために、いずれ

3) 日立製作所製 M680H

4) 同様のコンセプトに基づくもので、日経のミクロデータに関するものとしては豊原（1995）、日本開発銀行のデータに関するものとしては豊原（1999）がある。

5) 同様のシステムは神戸大学経済・経営研究所の安田，阿部による RIEB(1999) などがある。

のプラットフォームでも入手可能なフリーな言語である日本語版のgawk[6]を本体部分で採用した。

以上のことを背景にして、本稿では、われわれが開発した日本経済のマクロデータ抽出システムを紹介することを目的とする。まず第2章ではkg-edensの構造について、第3章ではデータ抽出の方法について述べた後で、第4章では今後の課題も含めたまとめを行う。

2. kg-edensの構造

経済のマクロデータを抽出するにはそのデータを特定するコード、期種（暦年，年度、半期、四半期、月次）、抽出開始期、末期、出力先を設定する必要がある。そのためのインターフェイスはgui[7]のようにノービスユーザに便利なものであればそれだけ一層システムへの負荷が大きく、さらにベテランユーザに対して不要な操作を強いることになり、逆にコマンドラインのみのインターフェイスであれば初心者への障壁を高めることになる。

そこで、われわれは次のようにシステムを2つに分割するという手法を採用した。具体的には図1にある本システムの構造のCとマークされた部分がもっぱら抽出を行うパートであり、ベテランユーザはワークステーションにログインしてコマンドラインからもっぱらgawkのスクリプトであるこの部分のみを起動することによってデータ抽出を行う。その際にはMTコードの検索といったサービスを受けることはできない。これはこのレベルのユーザであれば同じようなデータを期間や期種を変更して繰り返して抽出したり、以前の資料などを参照できると考えているためである。この場合には極めて軽快に動作することから、数百メガバイト程度の空き容量があるパソコンであれば、システムそのものを移行することも可能である[8]。さらに、unixやms-dosなどのコマンドラインでリダイレクト機能を活用すれば、あ

---

6) awk言語についてはAlfred V.Aho他(1989)を参照のこと。

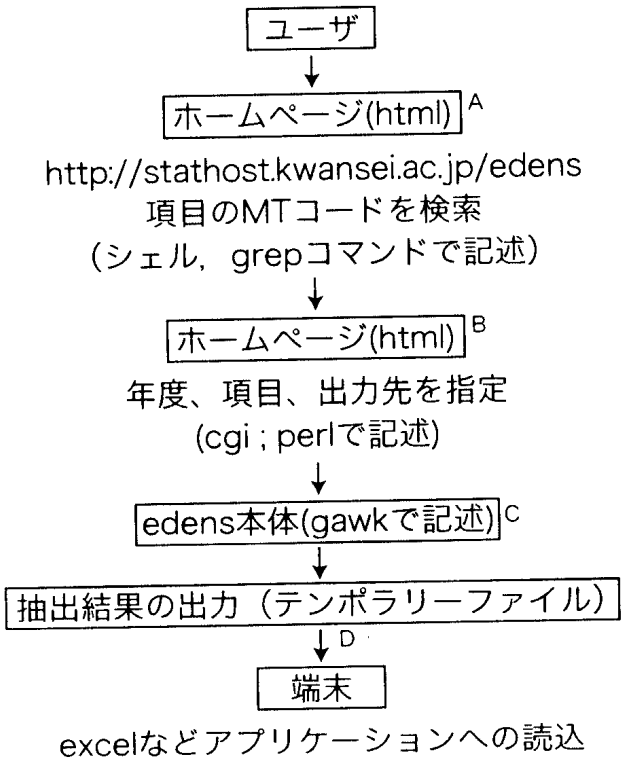
7) graphical user interfaceの略

8) 作動させるプラットフォームによってはディレクトリ構造が異なる場合があるので、その際には若干のスクリプトの変更が生じる場合もある。

らかじめ作成しておいたファイルに上述の各種パラメタを入力しておくことによって、毎回パラメタを入力することなく、類似した抽出操作を行うことができる。

ホームページを利用するノービスユーザの場合には、MT コード表が手元にない状況を想定しており、8 つまでの系列を同時に抽出できる。具体的には図 1 の A,B において、期種、抽出開始期と末期を選んだ後で、MT コードを検索するモードに入り、キーワードを入力するとそれにマッチする項目がリストアップされ、該当するものがあればその前にチェックボックスに印をつける。この処理は、unix の grep コマンドを使って各データのヘッダ部分を検索しているので、正規表現にも対応しているが、出力制御はしていないために入力する単語によっては画面に納まりきらない場合も考えられる。

図 1　kg-edens の処理フロー



このように、構造としては項目コードなどを設定する部分（図 1 の A,B）と実際にデータ抽出の処理を行う部分（図 1 の C）、さらには出力結果を転送する部分（図 1 の D）からなる。ユーザによって抽出されたファイルはテンポラリーファイルと

してサーバ上に作成される。これを画面に表示するか、端末に転送するかはユーザの選択による。

また、データ部分は717個のファイルからなり、MTコードごとに100の位ごとにまとめ、マイナス数値および欠損値の表現をパソコンなどで利用しやすい形式に変更した上でそれぞれにヘッダをつけたものを1レコードとしたものである。これは完全なシーケンシャルファイルであればアクセスに時間がかかり、また1コードごとにファイルを作製したのではファイル数が多くなりすぎることを考慮したためである。また、MTコード検索の際に用いられる、約3メガバイトのヘッダーリストファイルも別途用意されている。いずれも、日経データ社より購入した8mmテープのファイル（固定長形式で約92メガバイト）から必要なフィールドを抜き出して作成したものである。これらの作成にはperlを用いた。これは固定長ファイルの処理が容易であるためである。

3. データの抽出方法

まずノービスユーザの場合について述べる。

最初にAにおいてhttp://stathost.kwansei.ac.jp/edensにあるホームページにアクセスし（図2）、そこでBにおいて抽出するデータの期種（年次、半年、四半期、月次）、抽出期間を設定する（図3）。ここでは、年次の場合のみ暦年と年度の選択が必要となる。また元のデータより長い期種を設定することは可能であるがその逆の処理は行っていない。従って、元のデータが月次のときにはユーザはすべての期種を設定できるが、元データがたとえば暦年のときにはそれ以外の設定は行えず、その旨のエラーが表示されその処理はスキップされる。次に各データのMTコードを検索した後（図3）、項目名のMTコードを指定して抽出を実行する。

なお、上記の処理の結果は、データ部分とヘッダ部分に別れて出力される。このデータ部分は、たとえば先の場合では図5のような形式になっているために、excelなどの表計算ソフトで容易に読込むことができる。

次に、コマンドラインで利用した場合について述べる。メディア教育センターに

図 2　起動直後の kg-edens

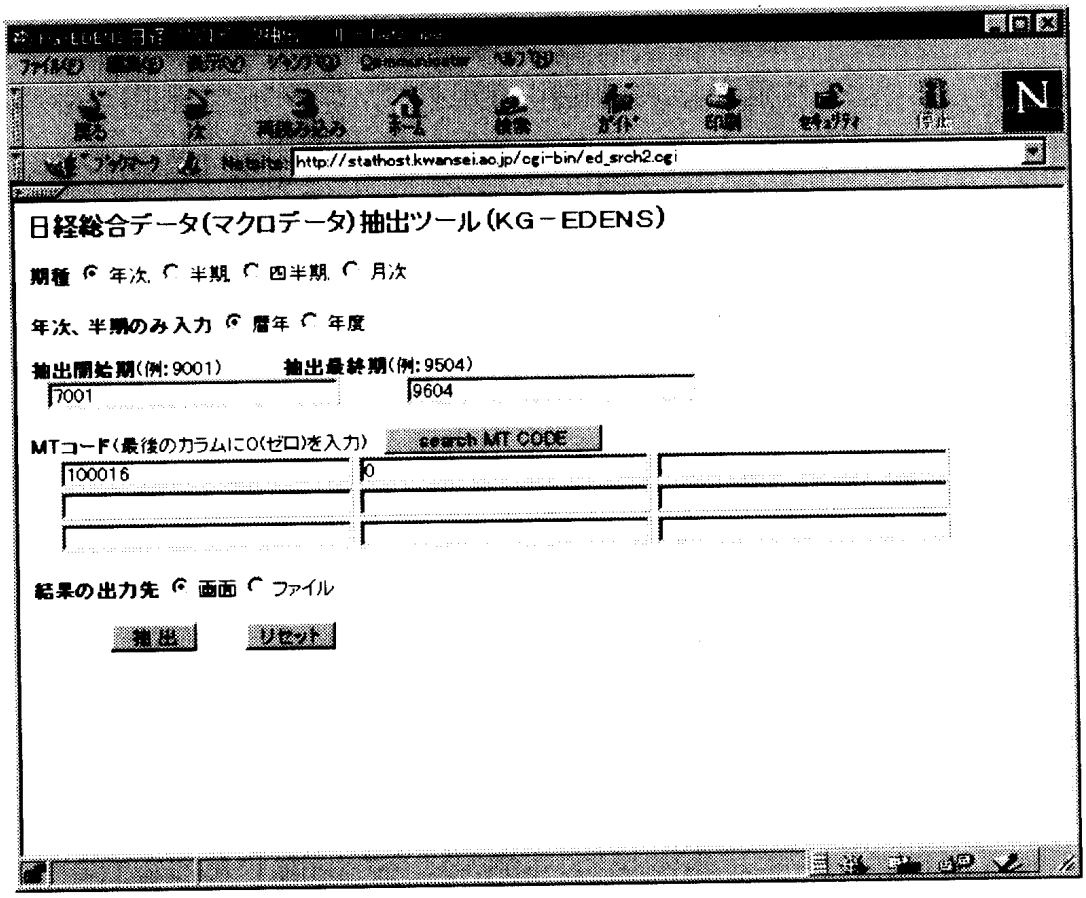


図 3　抽出の条件設定画面

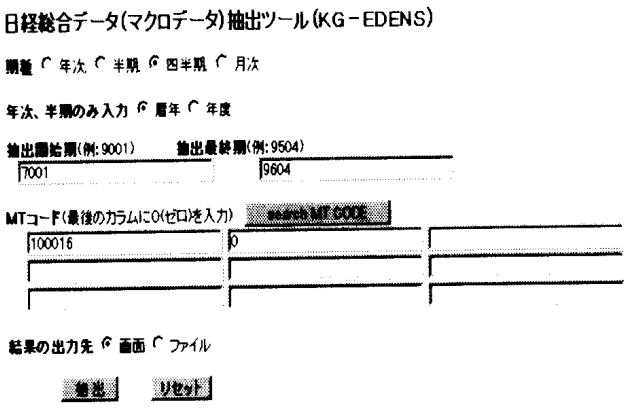


図４　コード検索の画面

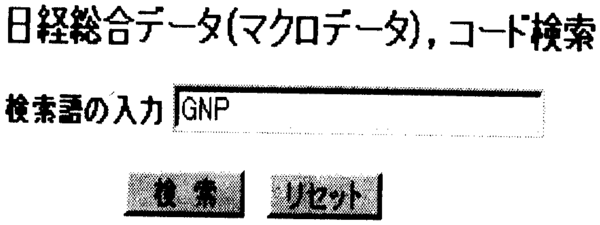


図５　kg-edens による出力形式

```
"period","00100016","","","",""
7001,16128.200000,0.000000,0.000000,0.000000,0.000000
7002,16792.510000,0.000000,0.000000,0.000000,0.000000
7003,18237.580000,0.000000,0.000000,0.000000,0.000000
7004,22030.070000,0.000000,0.000000,0.000000,0.000000
7101,18091.800000,0.000000,0.000000,0.000000,0.000000
7102,18627.380000,0.000000,0.000000,0.000000,0.000000
7103,19913.320000,0.000000,0.000000,0.000000,0.000000
7104,23959.380000,0.000000,0.000000,0.000000,0.000000
7201,20306.220000,0.000000,0.000000,0.000000,0.000000
7202,21063.840000,0.000000,0.000000,0.000000,0.000000
7203,22924.650000,0.000000,0.000000,0.000000,0.000000
7204,28106.130000,0.000000,0.000000,0.000000,0.000000
7301,24444.460000,0.000000,0.000000,0.000000,0.000000
7302,25729.400000,0.000000,0.000000,0.000000,0.000000
7303,27845.770000,0.000000,0.000000,0.000000,0.000000
7304,34499.830000,0.000000,0.000000,0.000000,0.000000
7401,28604.250000,0.000000,0.000000,0.000000,0.000000
7402,31052.340000,0.000000,0.000000,0.000000,0.000000
7403,33722.170000,0.000000,0.000000,0.000000,0.000000
7404,40618.060000,0.000000,0.000000,0.000000,0.000000
7501,32763.230000,0.000000,0.000000,0.000000,0.000000
7502,34693.420000,0.000000,0.000000,0.000000,0.000000
7503,36573.350000,0.000000,0.000000,0.000000,0.000000
7504,44139.870000,0.000000,0.000000,0.000000,0.000000
7601,36802.750000,0.000000,0.000000,0.000000,0.000000
7602,39214.370000,0.000000,0.000000,0.000000,0.000000
7603,41421.760000,0.000000,0.000000,0.000000,0.000000
7604,48978.010000,0.000000,0.000000,0.000000,0.000000
7701,41538.360000,0.000000,0.000000,0.000000,0.000000
7702,43803.760000,0.000000,0.000000,0.000000,0.000000
7703,45733.540000,0.000000,0.000000,0.000000,0.000000
7704,54454.440000,0.000000,0.000000,0.000000,0.000000
```

stathost の id 申請などが完了していることを前提とする。コマンドラインから

```
gawk -f /home/toyohara/data/edens06.awk /home/toyohara/dum
```

と入力する。これは gawk というソフトを、/home/toyohara/data というディレクトリにあらかじめ作成されている edens06.awk というスクリプトを読込んで実行するというものであり、これが先の C の部分に相当する。なお、最後の/home/toyohara/dum はダミーのファイル。

その結果、次のようなインタラクティブモードで kg-edens が実行される。流れ

図 6　コマンドラインによる実行

```
出力するファイル名：不要な場合は空行入力　：　testout E
期種を入力(1:年次, 2:半期, 4:四半期, 12:月次)　：　4 F
抽出開始期を入力(70年第1四半期なら7001)　：　7501 G
　　最終期を入力　：　9501 H
対象のMTコードを入力(0で終了)　：　100016 I
対象のMTコードを入力(0で終了)　：　0 J
```

について述べると、図 6 の E ではこの処理の結果を出力するファイル (testout) を設定している。ここでの名前に関するルールは unix に従うのでロングネームも可能ではあるが、端末に転送することを考えると、2 バイト文字の（日本語など）ではなく英数 8 文字以下にしておくことが望ましい。ここで空行入力 (何も入力せずにエンターキーを押すこと) すると、抽出結果は画面に表示される。ターミナルソフトによっては画面の履歴を取得できるのでそれをカットアンドベーストし、若干の修正を行えばそのままパソコンのアプリケーションに取り込むこともできる。

次に期種を指定する。これは年次、半年、四半期、月次の中から選ぶものであるが、年次を選んだときには暦年（1 月から 12 月）か年度（4 月から 3 月）かのいずれかを選択できる。

開始期、最終期については、現時点では西暦の下 2 桁と 2 桁の期内番号を設定する。具体的には年次データのときには 01 のみ，半期では 01 と 02、四半期では 01～04、月次では 01～12 といった数値が入る。年次のときも 01 が必要である。開始期と最終期が入れ代わっているときには何も出力されないので注意を要する。

最後に MT コードを入力する。このモードではあらかじめ抽出したい項目のコードは判明していることを前提にしているので、それをひたすら入力することになる。そして最後に 0 を入力すると、抽出を開始する。なおこのモードのときには抽出個数の制限はない。

最後に、コマンドラインでリダイレクトを利用する方法を述べる。先と同様に、

gawk のスクリプトを edens06.awk とし、読込ませるべきファイルを testin，出力ファイルを testout とするとき、あらかじめ図 7 のような testin を作っておくと、

```
gawk -f /home/toyohara/data/edens06.awk /home/toyohara/dum<testin
```

と入力しても同様の結果が得られる。さらにこの手続きそのものをシェルスクリプトとして登録し実行属性を与えておけば[9]，繰り返しが一層容易になる。

図 7　コマンドラインからの読込み用ファイル

```
testout
4
7001
9504
100016
0
```

なお、コマンドラインから MT コードなどを検索するときには、grep コマンドをもちいて直接ヘッダーファイル (gncent.hed) を調べればよい。具体的に GNP という単語を調べる場合について述べる。

```
grep GNP /stat/edens/data/gncent.hed
```

これは、grep というコマンドで gncent.hed というファイルの中にある GNP という語を含む行を表示せよというものである。なお、/stat/edens/data/は実際にこのヘッダファイルが置かれているディレクトリを指名している。従って、ユーザがパソコンにシステムを移植している場合には、この部分の記述は変更されねばならない。また、多数の表示が想定されるときには、さらに別のキーワードも追加できる。たとえば 90 年価格を示す 90Y を追加するには

9)　上記のシェルスクリプトを含んだファイルを goedens とするとき、以下のようにコマンドラインで入力する。

chmod +x goedens

```
grep GNP /stat/edens/data/gncent.hed | grep 90Y
```

とすればよい。なお、unix では大文字と小文字は厳密に区別されるので、その点の注意は必要である。また、ヘッダファイルには日本語のフィールドもあるので日本語による検索も可能ではあるが、端末からの日本語入力にはかなり煩瑣な作業が生じるのでここでは述べない。

以上のことから、コマンドラインで実行するためのパラメタ設定（MT コード検索・設定、期種、抽出開始期、末期の設定など）をノービスモードでは gui によって行っていることが明らかとなろう。

4. むすび

われわれのシステムを用いることによって，日経データ社のマクロ経済データからユーザの指定するデータを抽出できることになる．本システムの特徴としては、gawk という unix でも ms-dos でもフリーで利用可能な言語を利用していることに加えて，とくに，ベテランユーザではリダイレクトを用いることによって効率的に作業が行えるといった点があげられる。実際に抽出されたデータは csv[10] と言われるテキスト形式であることから、パソコンの表計算アプリケーション、計量分析に用いられる tsp,rats などにも簡単に取り込むことができる。

今後の課題としては，教育・研究活動をサポートする点からも収録データベースを増やすことに加えて、抽出そのものを容易にするための手法、つまりデータベースソフトを導入し、ODBC の活用によって excel などマイクロソフト社製アプリケーションソフトとの親和性も改善する必要があろう。

〈参考文献〉

豊原法彦，「ネットワークに対応した立命館大学経営データ教育システム;ws-RUBIES の開発」，『立命館経営学』第 34 巻第 4 号 67～88 ページ，1995 年 11 月 10 日発行.

豊原法彦，「DERFI-KG; Data Extraction for Ritsumeikan Financial Institute-

10) comma separeted variable の略

KaiGin data の構築」，『立命館経営学』第 37 巻第 5 号 125～154 ページ，1999 年 1 月 10 日発行.

安田　豊，阿部茂行,『RIEB データベースの研究』，神戸大学経済経営研究所刊行、研究叢書 52, 1999.

Alfred V.Aho、Brian W.Kernighan、Peter J.Weinberger （足立高徳 訳）,『プログラミング言語 AWK』，アジソンウェスレイ／トッパン，1989 年 11 月

Larry Wall and Randal L.Schwartz（近藤　嘉雪訳）,『Perl プログラミング』，ソフトバンク，1993 年 9 月発行