

Excelの文字列関数によるテキスト処理と 住所録データベース構築の演習

The exercise of text processing by the character string function of Excel and address book database construction

土橋 喜(愛知大学現代中国学部)

要旨

本稿は情報リテラシーの学習を終えた履修者を対象に、Excelの文字列関数を使った簡単なテキスト処理を行い、住所録データベースの構築を行うために開発した教材である。文科系や社会科学系の学生が、単なるデータベースの体験に留まらず、数千件程度のアンケート集計などに応用することを目標に、Excelを使い短時間にテキスト処理とデータベース構築を体験できる方法をまとめた。ウェブページに公開されている大学の住所録をダウンロードしてエクセルに張り付け、文字列関数を使って、住所録に必要な項目を抽出し、郵便の宛名印刷に必要なデータベースを構築する手順をまとめた。

キーワード：エクセル、文字列関数、テキスト処理、データベース、住所録

1. はじめに

本稿は情報リテラシーの学習を終えた履修者を対象に、Excelの文字列関数を使った簡単なテキスト処理を行い、住所録データベースの構築演習を行うために開発した教材である。文科系や社会科学系の学生が、Excelを使い短時間にテキスト処理とデータベース構築を体験できる方法を開発した。

現在では数多くのデータベースが作成されており、インターネットから利用できるものも多く、日常生活においても不可欠なものになっている。大学でもデー

タベースを扱う授業は、情報系の学部を中心に以前から実施されているが、ここではワードやエクセルなど情報リテラシーの学習を終えたレベルの学生に、データベースとはどのようなものを、1～2回の授業で体験的に学習させることを目標にしている。主に文科系や社会科学系の学生が対象であり、単なるデータベースの体験に留まらず、数百件から数千件程度のアンケート集計などに応用することを目標にしている。

また本稿で取り上げた教材は、エクセルを使った社会データ分析入門という授業で使っている。社会データ分析入門の

授業では、前半はおもに統計学の基礎理論を学習し、後半の3～4回を応用的な学習に充てている。毎時間エクセルを使って演習を行っており、本稿はそのなかの一部である。

データベースを構築し運用管理を行うDBMSには、大規模システム向けではOracle, MySQL, PostgreSQLなどがあり、小規模システム向けではMicrosoft社のAccessが多く使われている。しかし、これらのシステムを使って授業を行うには、前提となる知識が多々必要となり、半期または通年の授業時間を確保して行うのが一般的である。そのため簡単に試してみたい初心者にはハードルが高いと言わざるを得ない。

エクセルに備わっているデータベース機能を活用しても、データ管理、整列、検索、抽出、複数の小計の表示など、多面的な分析が行え、基本的なアンケート集計などには活用できる。加えて本稿が対象とする履修者は情報リテラシーでエクセルを学習している。そこでこれらの機能を使い、文科系や社会科学系の学生が、短時間に体験できる方法を考案した。授業で実施する際には、データベースの特徴を理解するため、授業時間で可能な限り多くのデータ(約900件)を扱えるように工夫して、データベースを構築するようにした。

以下は教材の内容であるが、履修者が自学自習できるように、解説およびエク

セルの操作手順と画面例を記載した。

2. データベース

データベース(database)は、大量のデータをコンピュータで迅速にかつ効率よく管理するために作成されたデータおよびそれらのデータを管理するソフトウェアの総称である。データベースという言葉は1950年代になって米軍によって初めて使われたもので、分散する資料(data)を1か所の基地(base)に集めて効率よく活用するという発想から生まれたと言われている。

データベースにはさまざまなデータを格納することができるのでその種類も多数ある。テキストを主体とする文献データベースをはじめ、音楽や画像を集めたデータベース、あるいは人口や景気変動などの数値データを集めたデータベースなどもある。個人で作成している住所録なども一種のデータベースである。

最近ではCD-ROMやDVDに収められたデータベースも多くなり、データベースそのものを持ち運べるなど扱い方が便利になっている。また従来から行われていた通信回線を使ったオンラインによるデータベースは、近年のインターネットの普及によって、ウェブページに広く公開されるようになり、データ転送のさらなる高速化と大容量化が進んでいるため、アクセスが容易になっている。



図1 「日本列島お国自慢」のホームページ

データベースに格納されたデータは、複数の利用者によって共有されて同時に利用されるのが一般的であり、そのための運用管理システムを備えている。データベースの操作や保守、あるいは管理するためのソフトウェアをDBMS (Data Base Management System)と呼んでいる。

最近ではデータの集まりを複数の表形式で表現するリレーショナルデータベースが中心となっている。また近年ではデータの集合を、手続きとデータを一体化してオブジェクトの集合として扱うオブジェクトデータベースと呼ばれる技術

が開発されており、大規模なデータベースで利用されている。

商用データベースには、新聞記事、雑誌記事、人物・企業情報、文献などの情報が検索できるものなどがある。近年インターネットのホームページの情報を収集したデータベースがいくつも作成されており、検索エンジンとともに社会的に重要な存在となっている。

利用者はキーワードやコマンドを入力して目的とする記事などを検索し、データをダウンロードしたりすることもできる。

最近ではインターネットから利用できる



図2 北海道の大学の住所録を表示させた例

ものが多くなり、プロバイダによっては国内外のさまざまな商用データベースへのゲートウェイサービスを行なっているものもある。

3. Excelのデータベース機能

Excelにはデータベース化した情報から目的のデータを検索したり、項目別に集計したりできる機能が用意されており、一般的に「データベース機能」ということがある。またデータベースから条件に一致するデータを検索できる「デー

タベース関数」も用意されている。Excelのデータベース機能を使うには、データをExcelがデータベースとして認識できるように、リスト形式で作成する必要がある。

ここではExcelに備わっている機能を使って、簡単なデータベースの作成とデータの検索などを取り上げる。

4. 住所録データベース

ここでは住所録を簡単なデータベースにすることを考えてみよう。インター

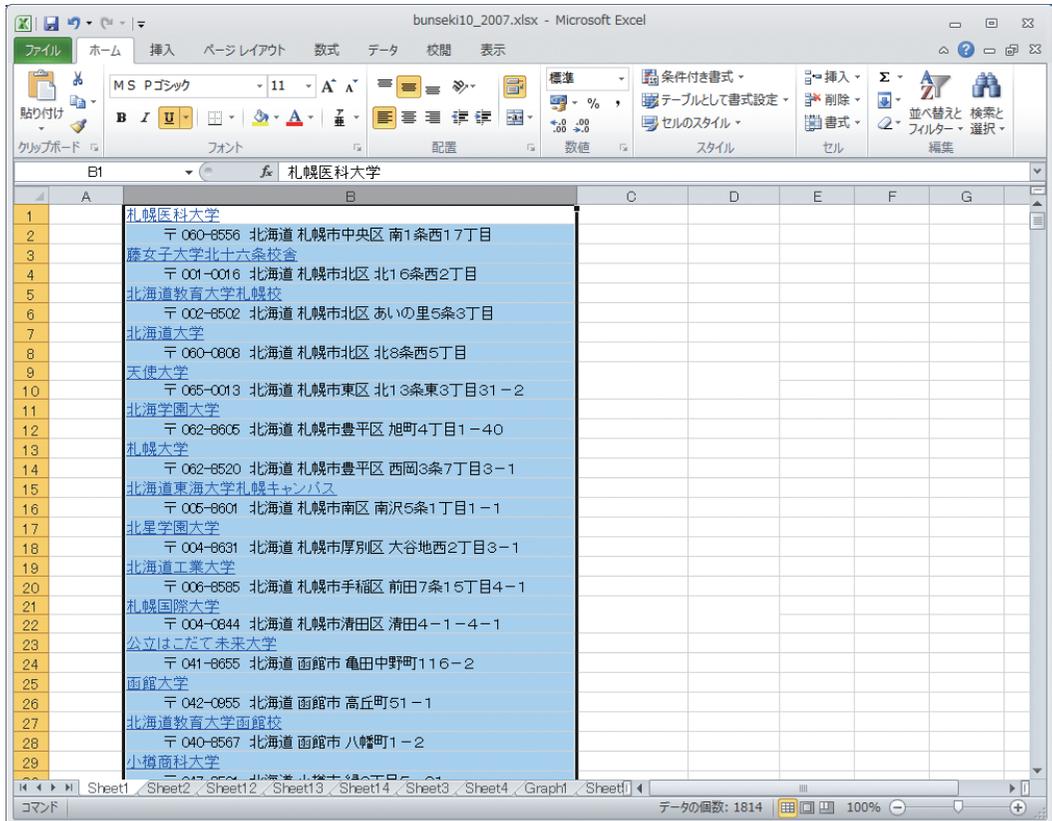


図3 貼り付け後の住所録データ

ネットに公開されている大学の住所録を使ってデータベースを作成し、そのデータベースを使ってどのようなことができるかを学習する。

インターネットに「日本列島お国自慢」(<http://www.towninf.co.jp/>)というページが公開されている(図1)。このWebサイトには日本各地の観光産業の情報を中心に、市区町村ごとに分類したさまざまな地域情報が紹介されている。その中に都道府県別に大学の住所録が作成されている(図2)。

大学の住所録を表示させるためには、

表示させたい都道府県の名前をクリックし、画面の左側にあるメニューの中から大学をクリックすればよい。例えば北海道にある大学の住所録を見たいときには、画面の中ほどにある「北海道」と表示してある部分をクリックし、次に画面左のメニューから「大学」をクリックする。

またはトップメニューの画面下部にある日本地図の形をした都道府県名一覧から、見たい県の名前をクリックする。さらに同様に画面左のメニューから「大学」をクリックする。

5. データの準備

以下で作成する住所録は著作権法により、外部に公開したり配布したりすることはできないため、以下ではこのことを念頭において進める。

先にExcelを起動して、新しいシートを開いておく。「日本列島お国自慢」のページで大学の名称・郵便番号・住所が表示された画面で、名称が表示してあるところの先頭から住所の終わり(例えば北海道ならその終わりまで)までをマウスの左ボタンを押しながらドラッグし選択する。次に選択した部分の上で右ボタンを押してコピーするか、編集メニューからコピーを実行する。

Excelをアクティブにして、B列の先頭から上でコピーしたデータを貼り付けるため、編集メニューから貼り付けを実行する。ひとつの都道府県が終わったら、次の都道府県を表示して同じようにコピーと貼り付けを繰り返す(図3)。なお名称・郵便番号・住所の行は列の先頭にだけあればよいので、最初に入力しておく。

ここではHTMLのリンクもそのまま貼り付けておく。そうすれば名称と住所がそれぞれワークシートの列ごとに分かれて貼り付けられるので、後々の作業に都合がよい。

またこの段階で「形式を選択して貼り付け」から「値」を指定して貼り付けを行うと、テキスト形式でデータが貼り付け

られるが、列に分割されないで1つのセルに張り付くため、何かの集計などに使いたいときは扱いにくい。

なおHTMLのリンクを削除してテキストデータで扱うと便利なこともあるので、データ全体をテキストにしたいときは、列ごとにデータを格納した後に、全体をコピーして空いている列または別のシートに、「形式を選択して貼り付け」から「値」を指定して貼り付けを行う。

6. 大学名と住所データの編集

全部の都道府県の大学住所録をコピーして貼り付けているとかなり時間がかかるので、ここでは北海道と東北地方ぐらゐまでのデータをExcelに貼り付けたことにして次に進めたい。

コピーしたデータによっては、部分的に形式が異なっていたり、データの一部が欠落したりする場合もある。そのようなときは形式の違う部分を修正したり、足りないデータを補ったりすることが必要になる。Excelの機能でうまく修正できないときは手作業で行う。

(1) 大学名の抽出

貼り付けた大学の名称と住所は2行でひと組のデータになっているが、Excelのピボットテーブルなどで集計したりするときは、リスト形式のデータになっているとやりやすい。そこで大学の名称と

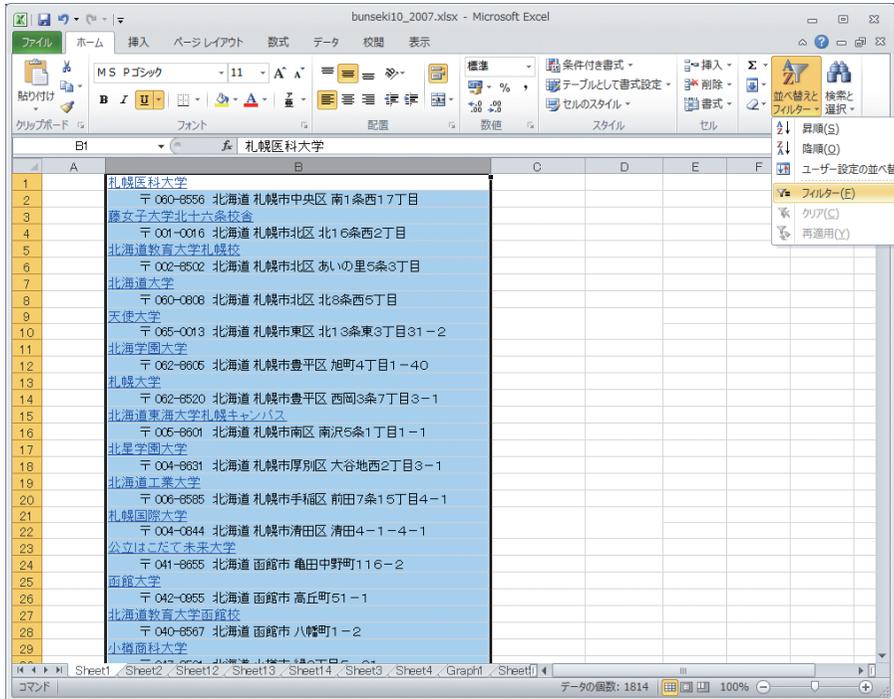


図4 フィルタの起動

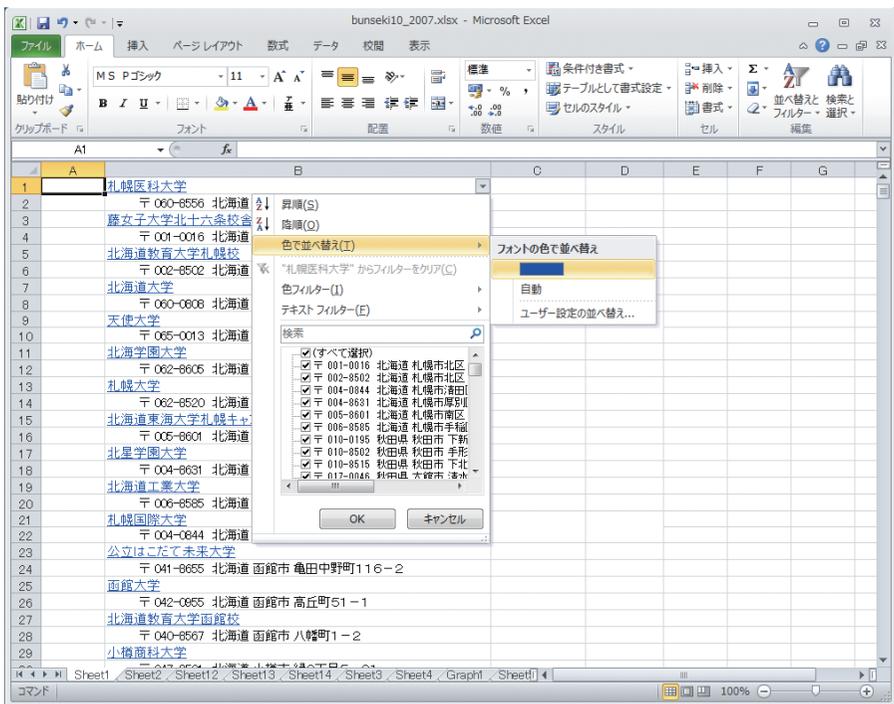


図5 フォントの色で並べ替え

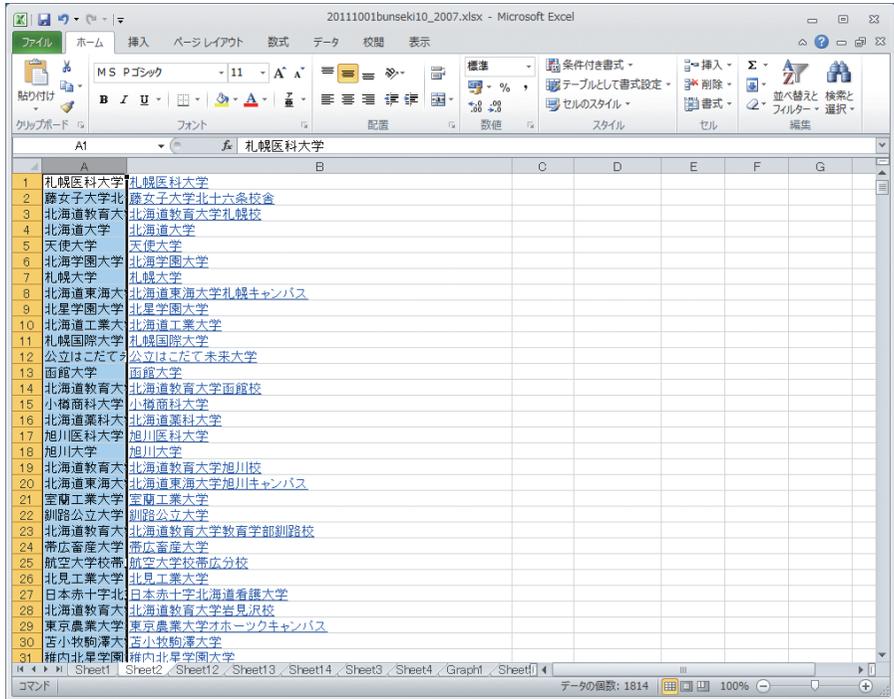


図6 名称データの貼り付け(値のみ)

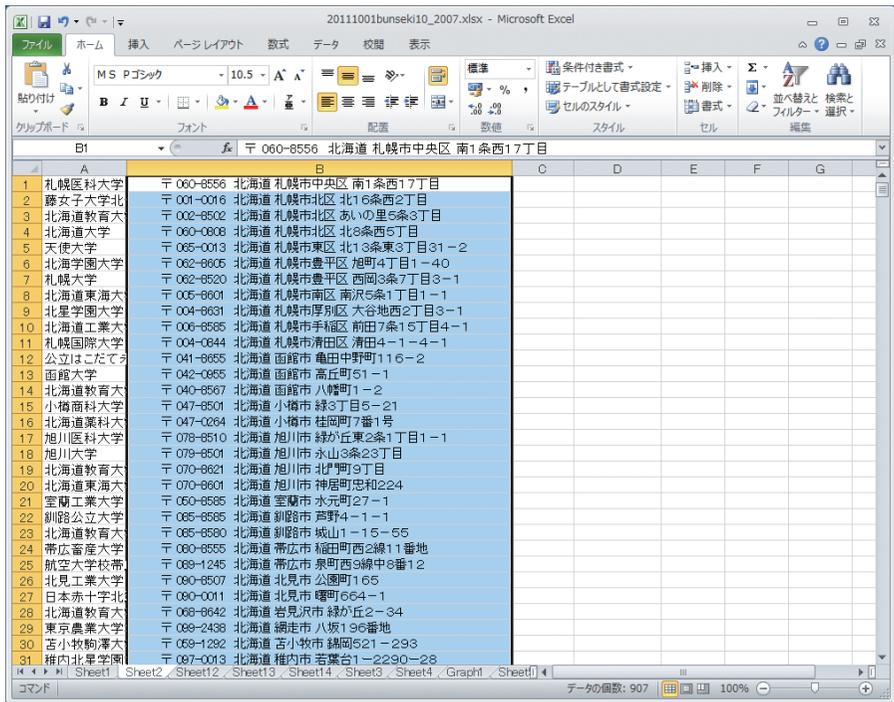


図7 名称データを削除して住所データを上方向にシフトした画面

住所をそれぞれ別々の列に格納するように工夫を行う。

◆まずA1のセルをクリックしておき、「ホーム」タブにある「並べ替えとフィルタ」をクリックして開き、「フィルタ」をクリックする(図4)。

◆次にB1セルに表示されたフィルタボタンをクリックして、「色で並べ替え」から青色のフォントで並べ替えを行う(図5)。

◆リンクが張ってある大学名は文字の色が青いので上位に並べられ、住所は文字の色が黒いので下部に並べられる。

◆大学名だけ別の列に抽出するため、大学名の部分を全部選択してコピーする。ここでは大学名をA列に作成することにして、コピーした大学名のデータをA1セルから貼り付けを行う。このときリンクは不要になるので、貼り付けるときは「形式を選択して貼り付け」から「値」をクリックして貼り付けを行う(図6)。

◆次にB列の大学名の部分を削除し、住所のデータだけを残す。このときに削除した後は住所のデータが上方向にシフトするように指定する。また大学名と住所のデータにずれがなく、1行に納まっているかどうか確認しておくようにする(図7)。

(2) 郵便番号の抽出

郵便番号と住所が一体のデータでは、宛名を作成するときに郵便番号を自動的

に入力することが難しいので、ここでは郵便番号と住所を分離してそれぞれ別の列に格納することにする。

ここでは文字列を操作するMID()関数を使う。MID()関数は文字列の指定した位置から、指定された数の文字を返す関数であり、半角と全角の区別なく一文字を1として数える。使い方の一般的な形式はMID(文字列, 開始位置, 文字数)となっている。MID関数を使い、住所を読み込んで郵便番号だけを抽出するため、ここではC1セルに以下の関数を指定する。以下の指定では、B1セルにある文字列を読み込み、3文字から開始して8文字(これは3+8で11文字まで)を抽出する(図8)。

=MID(B1, 3, 8)

(3) 住所の編集

郵便番号を抽出したあとは住所のデータの先頭から郵便番号を削除する処理を行う。先頭から11文字までは郵便番号のデータとスペースなので、住所データの位置は先頭から12文字目にあたる部分が都道府県名の開始位置になる。そのため全体の住所データの長さを計算し、それから郵便番号の部分の文字数を引けば住所だけの文字数を求めることができる。

Excelでは文字列の長さを求める関数にLEN()がある。LEN()は文字列の長さを文字数として返す関数である。バイト数に関係なく1文字を1として計算する。MID()とLEN()を組み合わせると以下のよ

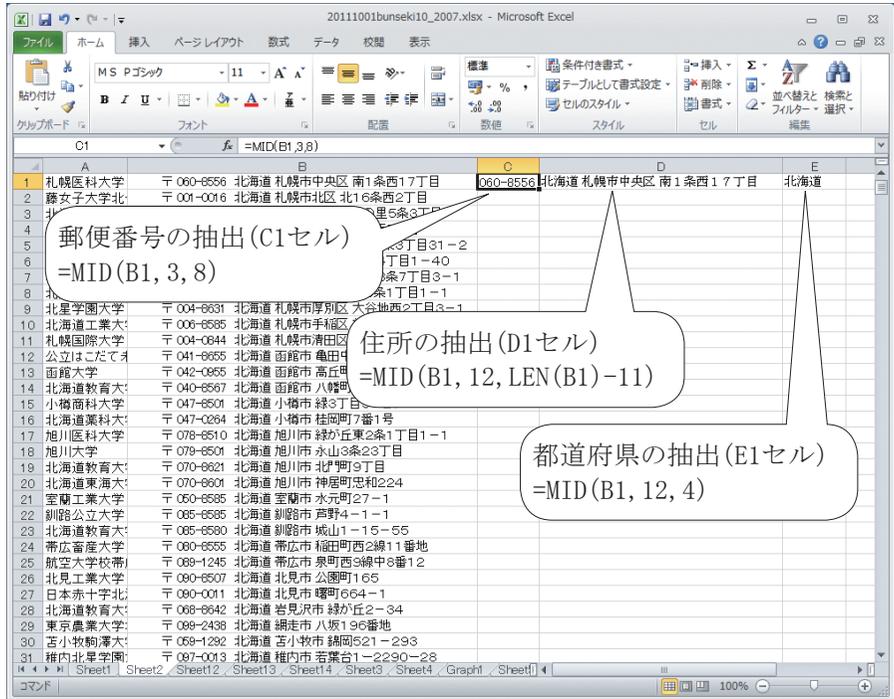


図8 郵便番号の抽出、住所の編集、都道府県の抽出

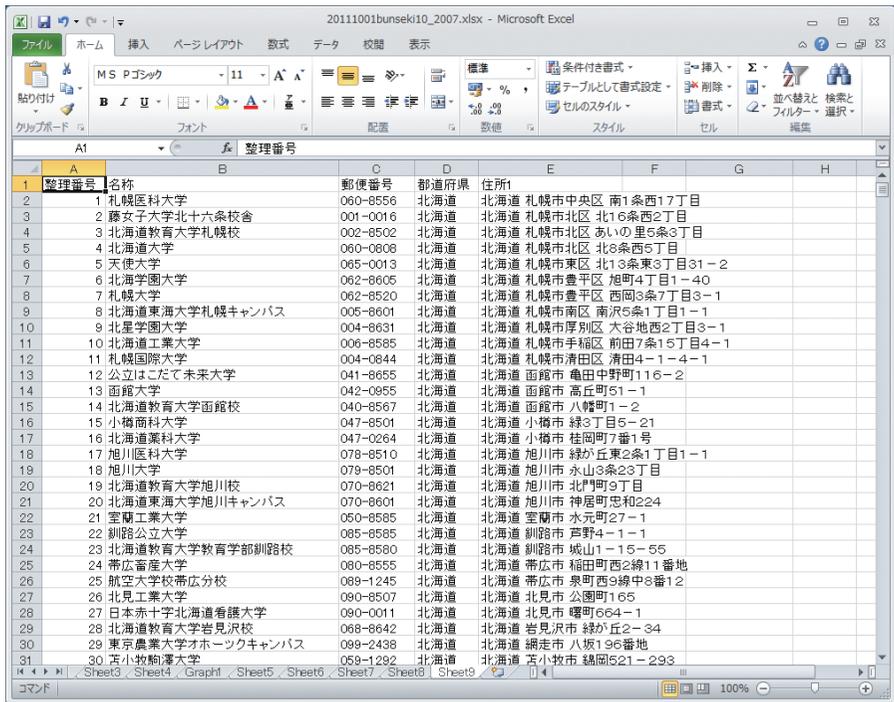


図9 貼り付けして表の体裁を整えた画面

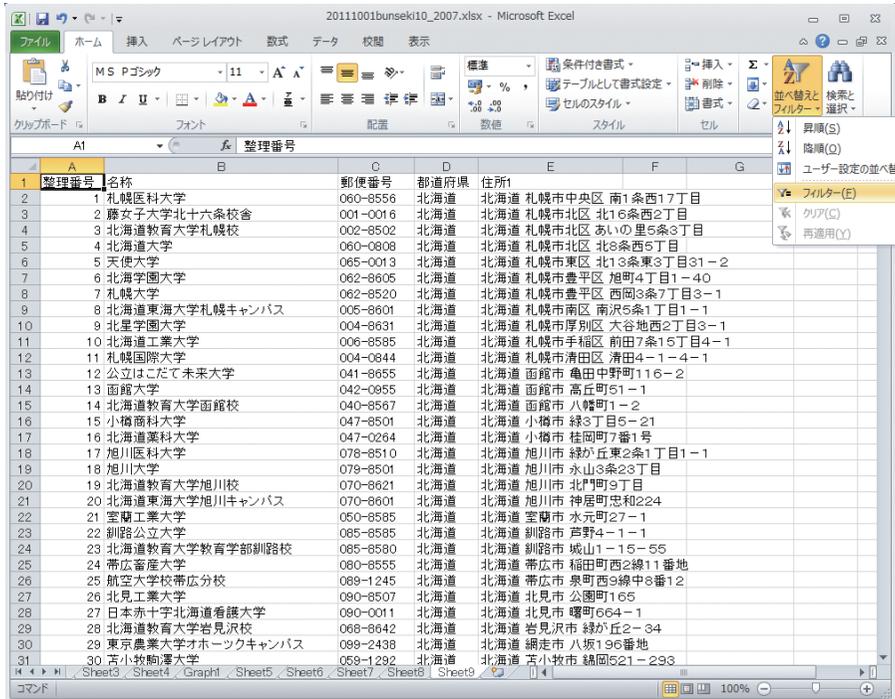


図10 フィルタの起動(その1)

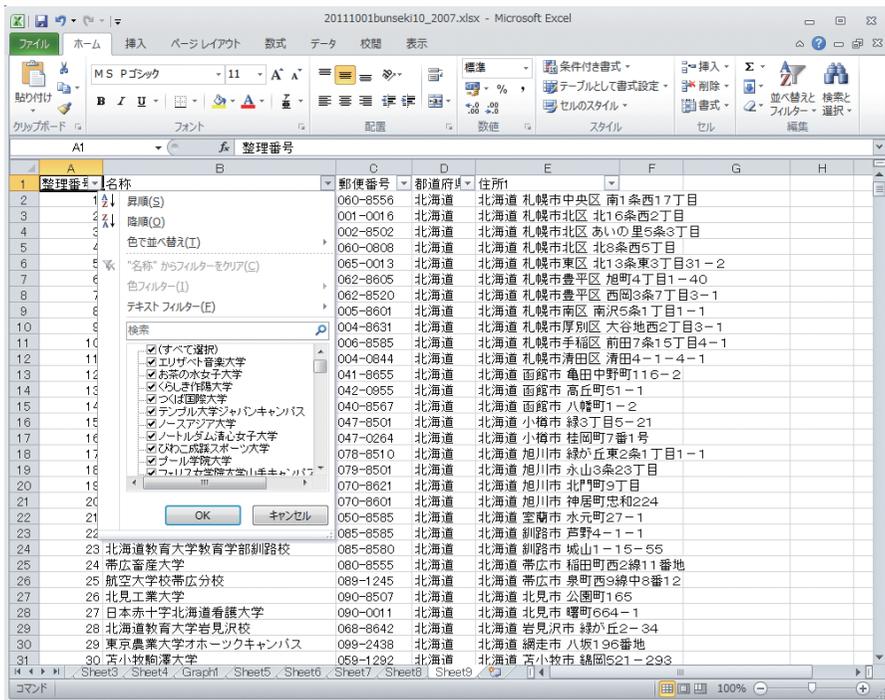


図11 フィルタの起動(その2)

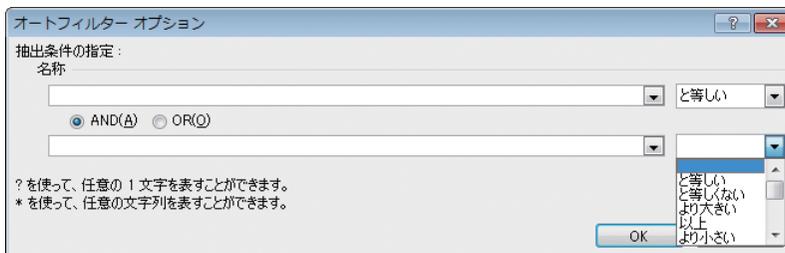


図12 オートフィルタのオプション

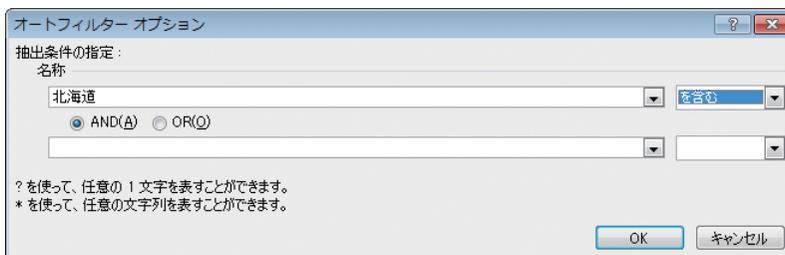


図13 オートフィルタオプションの設定画面

うにD1セルに関数を入力する(図8)。

=MID(B1, 12, LEN(B1)-11)

(4) 都道府県名の抽出

エクセルのピボットテーブルで都道府県別の大学キャンパス数を集計することを考え、住所のデータから都道府県名だけを抽出して、別のセルに格納する。都道府県名は3文字から4文字なので、郵便番号を含む住所のデータから、MID関数を使い先頭の12文字から開始して4文字を抽出する。以下のようにE1セルに関数を入力する(図8)。

=MID(B1, 12, 4)

7. データの検索

Excelにはシートに格納されたデータ

を検索する機能が備わっている。「フィルタ」はデータベース内の大量のデータから、手軽な操作で必要なデータを取り出してくれる機能である。フィルタを使うときは、「ホーム」タブの「編集」にある「並べ替えとフィルタ」をクリックし、「フィルタ」を選択して起動する(図10)。

フィルタが起動すると、セルの右端に三角の矢印が付いたボタンが表示される。そのボタンをクリックするとフィルタの機能を使うことができる。フィルタの主な機能には次のようなものがある(図11)。

(1) すべて

「すべて」をクリックすると全部のデータを表示する。

(2) 昇順, 降順, 色で並べ替え

昇順, 降順に並べ替えて表示する。ま

た文字などの色で並べ替えができる。

(3) テキストフィルタ

テキストフィルタをクリックすると「オートフィルタオプション」が起動し、いろいろな検索オプションを指定することができる(図12)。

例えば「北海道」という文字を含むデータを検索したいときは、名称の欄に「北海道」を入れ、右の選択肢から「を含む」という表示を選択し、OKボタンを押すと検索が実行される(図13)。

検索結果の表示を元に戻したいときは、オートフィルタのメニューから「すべて」をクリックする。

8. ワイルドカード文字と検索

ワイルドカードは該当する文字が分からないときや、複数の表記が使われており、それらを一括して検索したいときなどに使われる。オートフィルタオプションのワイルドカードには、「?」(疑問符)と「*」(アスタリスク)がある。これらのワイルドカードに使われる記号はいずれも半角文字で入力する(図13)。

(1)? (疑問符)

任意の1文字を?で指定するときは、例えば「インターフェ?s」と入力すると「インターフェイス」や「インターフェース」が検索される。あるいは「北海道??大学」と入力すると、「北海道工業大学」や

「北海道薬科大学」などが検索される。

(2)* (アスタリスク)

任意の数の文字を*で指定するときは、例えば「北海*」と入力すると“北海学園”や“北海道大学”などが検索される。

9. 都道府県別大学キャンパス数

ここではピボットテーブルを使って、入力したデータから都道府県別の大学のキャンパス数を求めてみよう。なお学校法人の数とは異なり、ひとつの学校法人が複数のキャンパスを持っていることに注意する。

(1) まずD列の都道府県名から計算するので、D1のセルからデータの最後までドラッグして選択する(D列全体を選択しても計算されるが、その場合はピボットテーブルの最後に空白行が表示され、グラフにも最後に空白の表示がなされる。しかし大きな問題はない)。

(2) ここではピボットテーブルとピボットグラフの両方を作成するため、「挿入」タブの「テーブル」から「ピボットテーブル」をクリックし、「ピボットグラフ」を選択して起動する(図14)。

(3) 「ピボットグラフ」を選択すると、「ピボットグラフ付きピボットテーブル

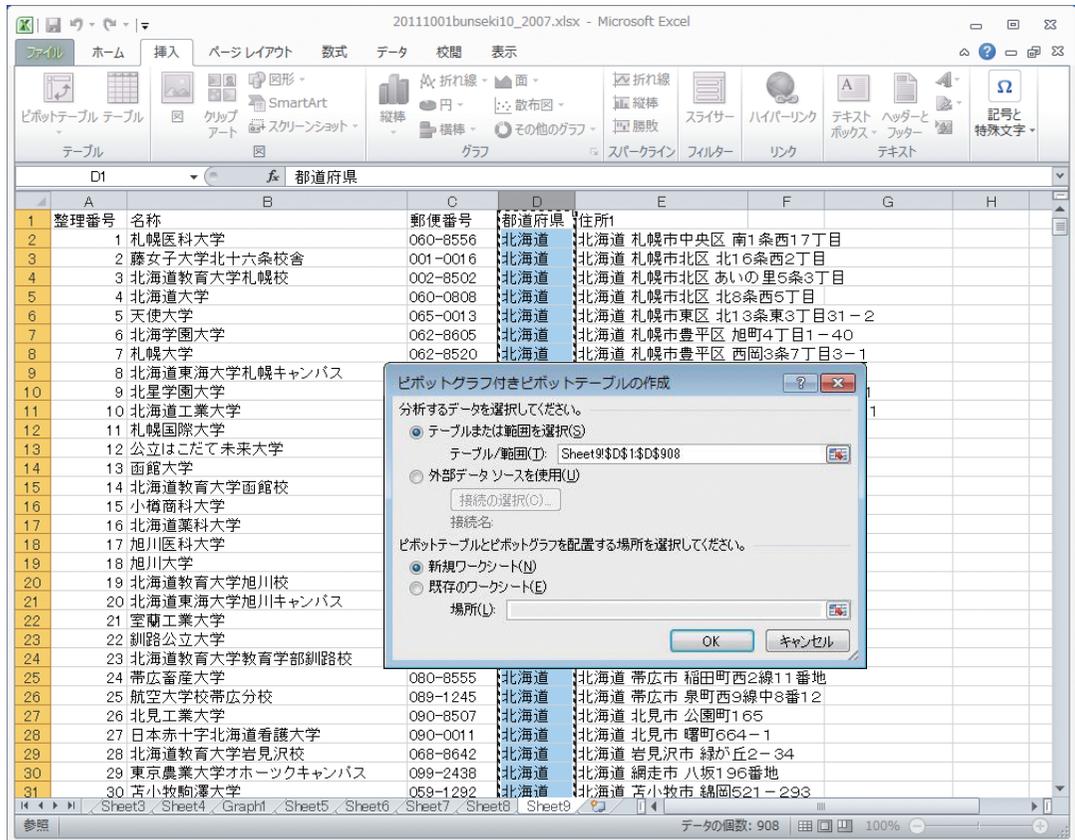


図14 ピボットグラフの設定画面

の作成」が開くので、「テーブルまたは範囲を選択」と「新規ワークシート」をチェックする(図14)。

(4) 「OK」をクリックすると新しいワークシートに、グラフエリア、ピボットテーブル、ピボットグラフフィルタ、ピボットグラフのフィールドリストなどのピボットグラフツールが開く(図15)。

(5) 次にピボットテーブルを作成するために、「ピボットテーブルのフィー

ルドリスト」に表示された都道府県名のチェックボックスをクリックする。これによってワークシートに行ラベルとデータが表示され、軸フィールドにも都道府県名が表示される(図16)。

(6) 都道府県別のデータを集計するため、「ピボットテーブルのフィールドリスト」に表示された都道府県名を「Σ値」のフィールドへドラッグする。「Σ値」のフィールドにはデータの個数という見出しが表示される(図17)。

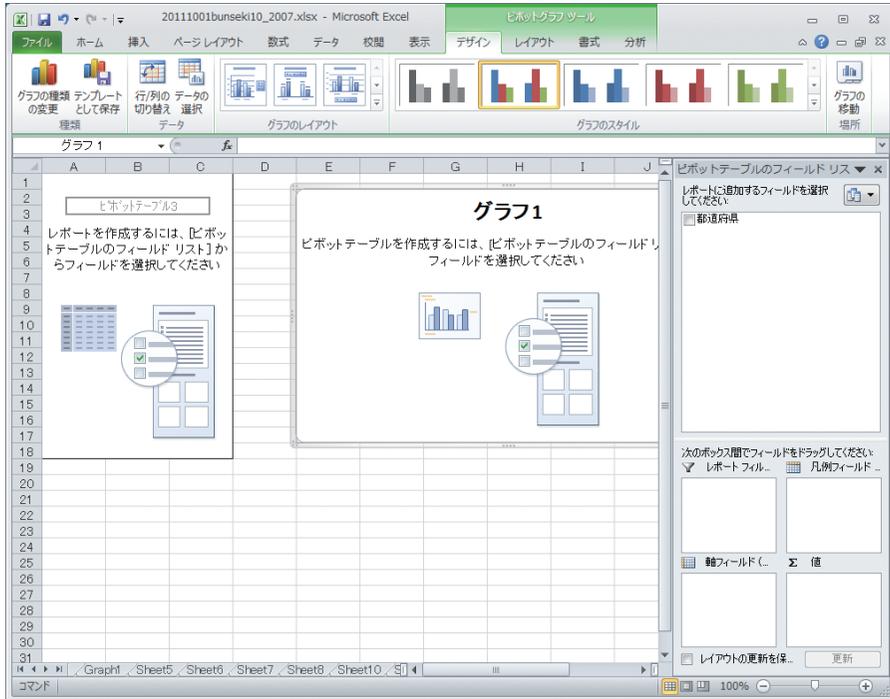


図15 ピボットグラフツールの表示

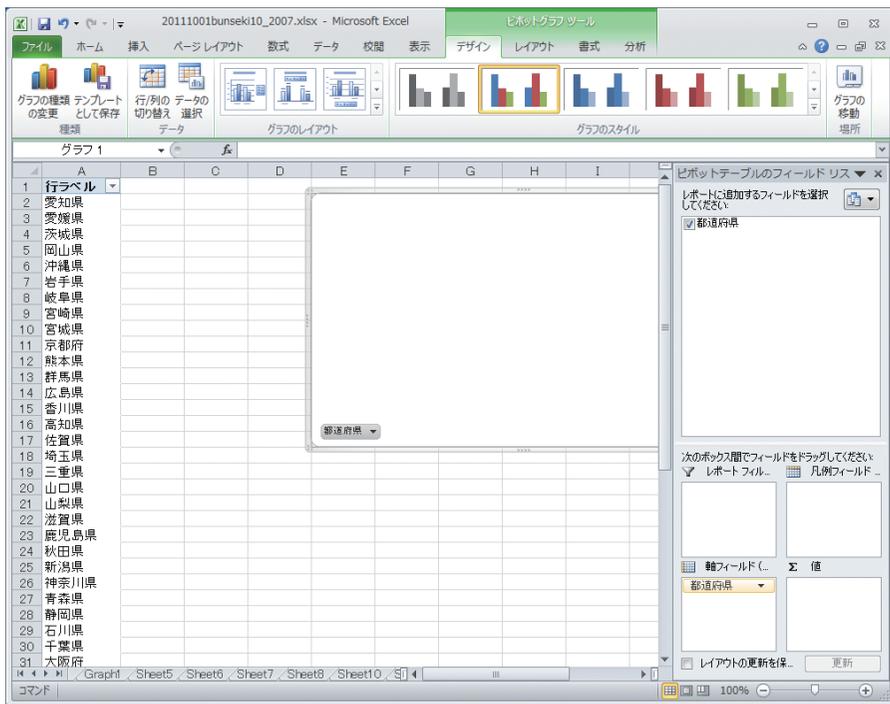


図16 都道府県名をクリックした後の行ラベルと軸フィールド

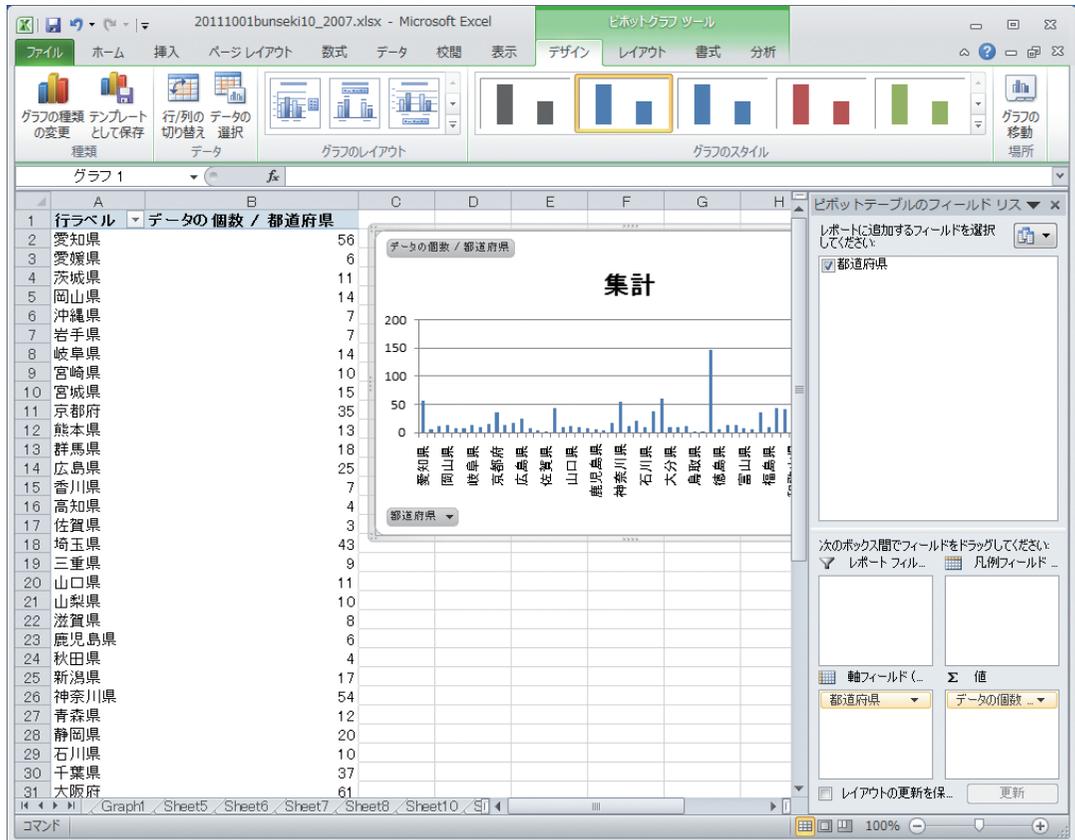


図17 ピボットグラフ作成の画面

(7) 以上の操作でピボットグラフとピボットテーブルが作成される。なおピボットグラフは特別な操作をしなくてもピボットテーブルと同時に作成される。必要に応じてグラフの形式や大きさなどの調整を行い、タイトルや凡例を適切に入力する(図17)。

10. 日本地図への表示

本稿の内容には都道府県名が含まれているので、地理情報システム(GIS)のデータとしても使えることから、エクセルとGISの連携した活用手法を取り上げていく計画である。例えば、MANDARAは地理情報システムの一つであり、上記のエクセルの数値データに、MAP、TITLE、UNITの3つのタグを追加するだけで、エクセルのデータを読み込み、日本地図上に数字のデータをグラフ化して表示することができる²⁾。

11. まとめ

以上の内容は今学期の授業で実施した直後のため、履修者の評価などは今後の

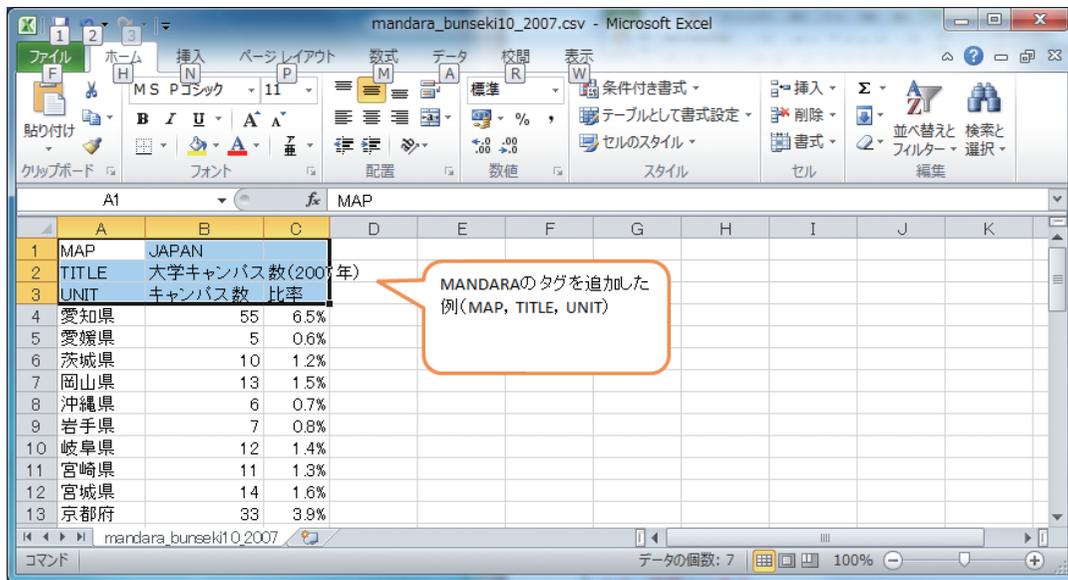


図18 エクセルのファイルにMANDARAのタグを追加した例

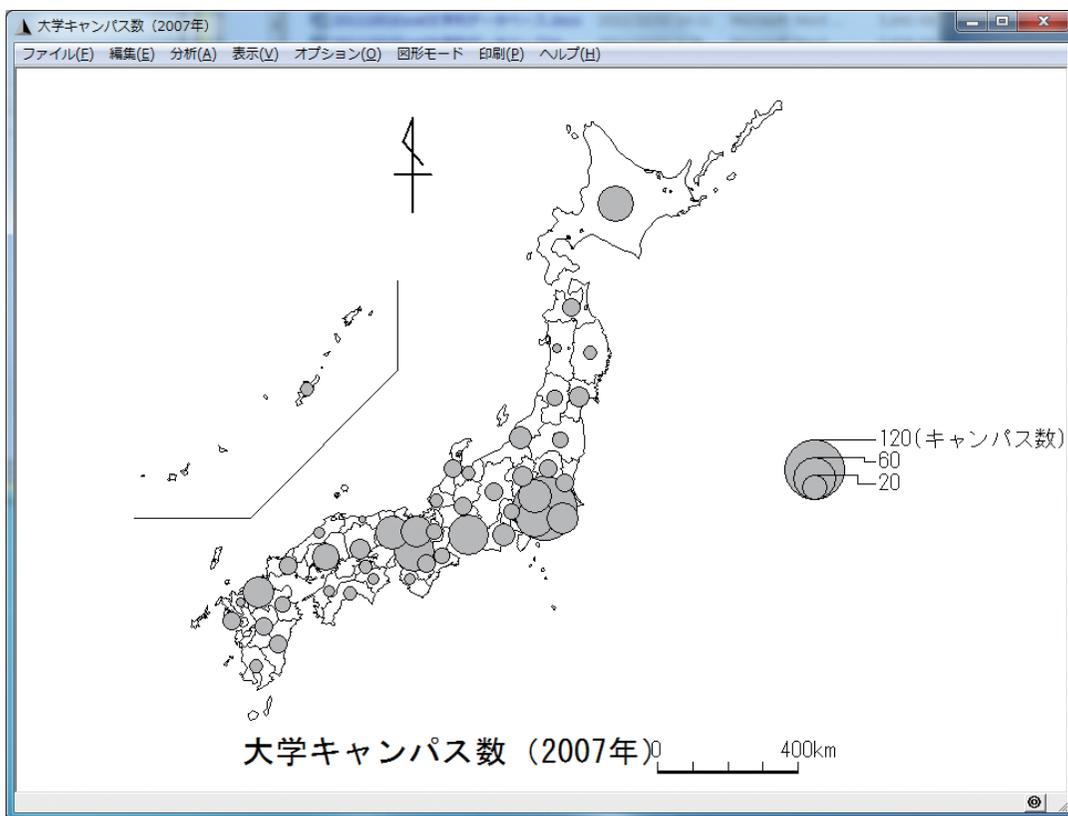


図19 MANDARAによる描画の例

課題である。しかし、2004年にデータの処理方法に文字列関数を使わないで、ウェブページからデータを貼り付けるだけで、住所録データベースの構築を授業で取り上げたことがあったが、履修者の反応はまずまずであった。履修者によってデータベースの構築に時間がかかり、全部のデータを入力する履修者がいる一方で、数十件の入力に留まっている履修者が少なからずいる状況であった。しかし、今学期は以前と比較してほとんどの履修者が全部のデータを入力し、データベースを完成していた。

文献

- 1) 土橋喜:Excelによるデータ分析の基礎, 愛知大学情報処理センター紀要「COM」, Vol. 13, No. 2(第23号), pp. 47-129 (2003).
- 2) 谷謙二著:MANDARAパーフェクトマスター, 古今書院, pp. 330, 2007.
URL:<http://ktgis.net/mandara/>