北海道植物データ処理システムの開発について（その1）

日野間　彰

（研究の概要）
北海道の植物と植物群落に関する各種データの保存と効率的かつ正確なデータ処理を可能とさせるために、主として道内各地の植物社会学的群落調査データと植物相データを対象としてデータベースを設計・作成した。その結果、道内の植物や植物群落について、大量のデータのデータ処理が可能となり、北海道植物データ処理システム（略称ＥＣＰＬＡＮＴ）の開発によって、高等植物の道内分布図の作成、森林群落のクラスター解析、特定の植物種や植物群落についてのデータ検索が可能となった。

1．研究の目的

北海道開拓の歴史もすでに百年をゆうにこえ、19世紀後半から数多くの研究者によって調査され報告されてきた植物に関する知見も数多く蓄積されてきている。さらに近年においては、開発計画に伴う環境アセスメントの実施も数多くなされてきており、北海道の植物に関するデータ量も膨大となってきている。それにもかかわらず、北海道のこれらの分野における科学的・技術的進歩は、国内の他の都府県に比べて、残念ながらまだ後進の位置に甘んじているといえる。

本研究は、広大で且つ豊かな自然を擁するこの北海道でこそ植物をはじめとする自然環境の第一級の研究と科学技術の発展が必要であるとの認識のもとに、その第一歩として、北海道の植物環境を定量的に把握することのできるデータベースを作成しその情報処理システムを実験的に作成することを目的として行ったものである。

2．研究の対象とするデータの範囲

本研究では、北海道の植物群落に関して調査・研究されたあらゆる形の静的なデータ（資料）および北海道の植物分布に関するあらゆる形のデータを対象として処理できるシステムをめざした。

実際に研究の対象として用いた文献・資料の数は延べ517件（1990年９月現在）で、最も古いものは1880年の標本に関する資料である。対象として用いた文献・資料の数をデータ区分名および年代ごとにとりまとめると、表１のとおりである。なお、データ区分名とは当該データのデータ形式の区分基準に基ずく名称であり、データ形式の区分基準は、後述するフィルドデータのフォーマットの検討の過程で作られた基準である。

表１　年代別・データ区分別文献・資料数

年代	データ区分名			計
	方形区データ	帯状区データ	分布データ	
1880	0	0	8	8
1890	0	0	17	17
1900	0	0	8	8
1910	0	0	17	17
1920	0	0	26	26
1930	0	0	43	43
1940	0	0	20	20
1950	1	9	30	40
1960	6	12	22	40
1970	71	13	49	133
1980	78	4	83	165
合計	156	38	323	517

（注）・1990.9現在
・標本データは分布データに含めた

3．研究の方法

本研究の内容は大きく分けて、
①データベースの作成
②データ処理システムの作成

の２つの作業からなる。ただしこの２つの作業は全く別々に行われるわけではなく、システムを考えながら最も望ましいデータベースの形を検討し、データベースの形を考えながらシステムを作成していくというように、つねに他方の作業を眺めながら進められていくものである。したがって、これらは時間的にも平行して作業が進められた。研究全体の作業フレームは図１のとおりである。

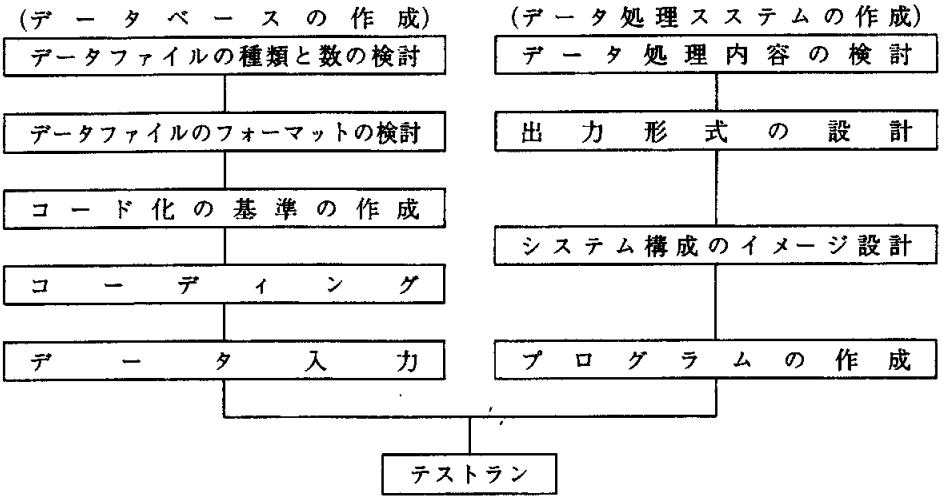

図１　全体の作業フレーム

データベースの作成は、文献・資料のデータを統一された内容と形式（フォーマット）に則ってデータ化しデータファイルとして登録・保存する作業であり、そのためにはフォーマットの設計、コード化の基準の作成などの手続きが必要である。またデータファイルも何種類かを用意する必要があり、どのようなデータファイルを用意すべきかを検討することも重要な手続きである。

データ処理システムの作成は、データベースを使って必要な情報へと加工処理するシステム作ることであり、システムは複数の段階のプログラムから形成される。ここではシステム全体のイメージを設計するとともに、どのような出力（計算結果）が必要であるかを検討することが主要なテーマとなる。

４．データベースの作成

(1)データファイルの種類と数の検討

データを加工処理して手にしたいアウトプットを得るためには、まず処理対象とするデータが整備されていなければならない。このようなデータを総称してデータベースとよんでいるが、本研究で取り扱うデータベースの場合には、「新知見などによってもデータそのものの価値が変わらないこと」、「日付、場所、調査方法などの基本的事項のもれがないこと」、「誤入力や原典のミスなどによるの誤りを確認し訂正することが可能であること」が重要な視点であると考え、データファイルの内容の設計にあたっては以下の条件を満たすよう配慮した。

①できるだけ加工されていない原データをそのまま使用すること

②どのような処理をする場合にも必要最低限のアイテムを網羅していること

③文献や資料の原典を探索することが可能であること

技術的には必要な全てのデータを１個のデータファイルとして作成することも可能であるが、データの作成・修正（入力）、データの保存・維持、データ処理のしやすさなどの観点から、一般に行われているように、データの内容に応じて幾つかのファイルに分けて作成するのが望ましい。本研究の場合、その目的を達するためには少なくとも以下の４種類のデータファイルの作成が必要と判断した。

①フィルドデータファイル
②植物分類データファイル
③文献データファイル
④処理条件設定データファイル

フィルドデータファイルは、文献・資料に記載されているデータを、コード化の基準に従って入力したデータファイルであり、各調査地点ごとに調査地点番号、調査時期、調査場所、調査者、立地の記録、階層、出現植物種類名、優占度などの細かな記録をデジタルデータ化しファイルたものである。フィルドデータの価値が将来的にも変わらないようにするため、フィルドデータファイルにはできるだけ加工されていない原データをそのまま使用して作成できるよう、コード化の基準の作成などに十分に配慮する必要がある。

植物分類データファイルは、フィルドデータファイルに登録されているすべての植物分類単位（植物種類）の和名、学名、分類上の位置などについてデータ化したもので、コード化されて保存されている植物の和名や学名を照合するために必要であるとともに、複数の和名や学名を有する同一の分類単位の判定にも必要である。また、このデータファイルによって一定のヒエラルキーに基づいた植物分類リストの作成が容易となるほか、各分類単位の特性（生活型、貴重性など）を加えることにより、例えば貴重な植物だけを検索するなど、特性に応じたデータの加工・処理が可能となる。

文献データファイルは、フィルドデータファイルに登録されているすべてのデータの出典（文献・資料）に関するデータファイルで、文献名、資料名、著者、発行年月日、発行者名などがデータ化されて保存される。このファイルにより文献や資料の原典を探索することが可能となる。

処理条件設定データファイルは、検索の条件、演算の内容の指定、アウトプットの内容の指定など、データ処理の条件を指定するデータファイルで、ここで指定された条件に基づいてコンピュータ処理がなされる。

基本的には以上の4種類のデータファイルがデータベースを形成する。データファイルの数も4個でよいわけであるが、フィルドデータファイルについてはデータ入力時における機械操作ミスによるデータの破壊を防ぐため、また植物分類データファイルについては1レコードに記録できるデータ容量の関係から、それぞれ3つと2つのデータファイルに分割したため合計7つのデータファイルが作成された（表2）。

実際にはこの他に、データ処理の過程で必要となる各種のデータファイル（コントロールデータファイル）があり、これらはデータ処理の目的に応じて随時作成される。処理条件設定データファイルもコントロールデータファイルのひとつと見做すことができる。また広義の解釈で考えればプログラムそのものあるいは演算の途中で生成される数多くの作業ファイルもデータファイルの一つとみなせるが、説明が繁雑になるおそれがあるためここではデータファイルとして取り扱わない。

表2　ＥＣＰＬＡＮＴのデータファイル構成

フィルドデータファイル	入力作業用フィルドデータファイル 一時運用フィルドデータファイル 永久保存用フィルドデータファイル
植物分類データファイル	分類・和名データファイル 学名データファイル
文献データファイル	文献データファイル
処理条件設定データファイル	処理条件設定データファイル

(2)データファイルのフォーマットの検討

①フィルドデータファイルのデータ形式の区分

各種の植物調査資料をフィルドデータファイルとして入力するためのフォーマット（様式）を検討する前に、資料となるデータを一定の基準によっていくつかの形式に分類して各形式ごとのデータの特性を整理して把握しておく必要がある。

北海道の植物群落および植物分布に関するデータの種類を整理してみるとおおむね次のいずれかに区分できる。

・北海道内で調査された植物群落調査資料
・北海道内の特定の地域あるいは北海道全体の総括的な植物目録
・北海道内あるいは北海道を含む地域の植物図鑑または植物写真集
・特定の植物種あるいは分類群の分布に関する資料で北海道に関係するもの
・北海道内で採集された植物標本およびそれを含む植物目録

これらのうち植物群落調査資料に含まれているデータの形式とそれ以外の資料に含まれているデータの形式には明瞭な違いがある。

植物群落調査資料に含まれているデータはほとんどの場合、調査方法により方形区法か帯状区法かのどちらかに分けられる。そしていずれの調査法による場合でも基本的に調査日、調査者名、調査地点の位置、調査面積、傾斜、海抜、階層、出現植物名およびその量的尺度などが調査地点ごとに調査され報告される。これに対し植物群落調査資料以外の資料に含まれているデータはほとんどの場合、該当する植物の確認・採集日、確認・採集者名、植物名、同定者名が報告される程度であり、特定の地域に関するデータ量としては前者の方が後者に比べ圧倒的に多いのが普通である。

そこで本研究ではデータベース化するフィルドデータを、そのデータの種類とデータに含まれているデータの項目（以下アイテムという）の差異を勘案して表3に示すように4つに区分した。

表3　フィルドデータのデータ形式の区分基準

データ区分名	データの種類・内容
方形区データ	方形区法で調査され報告されている植物群落組成調査資料
帯状区データ	帯状区法で調査され報告されている植物群落組成調査資料
分布データ	植物目録、植物産地報告、その他の植物分布記載
標本データ	標本、標本目録

上記の議論から容易に理解されようように、植物群落組成調査資料については1調査地点での調査を単位としてデータファイル化するのが妥当である。以下この単位を調査スタンドと呼ぶこととする。分布データと標本データについても、同一の日時に同一の場所（地域）で同一の調査者によって同一の調査方法により調査された調査資料については原則的に一つの調査スタンドとして取り扱うこととした。

②フィルドデータの識別コードの検討

次に、各々のフィルドデータがどの文献から引用されているのかという情報を保存するために必要な識別コードの検討を行った。この識別コードにより、フィルドデータファイルと文献データファイルとの情報の接続が可能となる。

識別コードとして考えられる最初の案は、ひとつひとつの文献・資料ごとに順次に番号を割り当てていく方法であるが、文献・資料の入手の時期の問題あるいは報告された資料の調査時期と発表時期とのタイムラグの問題などによりデータ入力の順番が必ずしも調査の順番にならない場合が多く、データ管理が繁雑になり間違いを生じさせる原因ともなりうる。そこで、識別コードの最初の4桁を調査年代（西

暦）に割り当て、識別コードのみで調査の時期がわかるようにすることとした。さらに各年代ごとに文献・資料を識別するための番号（以下文献番号という）を付すことにより、調査年代と文献番号の組み合わせによって文献・資料が特定できようにした。文献番号の桁数については、1年間に報告されるであろう文献・資料の量を勘案して3桁の数字とした。これにより文献・資料を特定するための識別コードは次のとおり全部で7桁の数字で表すこととなる。

“文献・資料を特定する識別コード”＝“調査年代”＋“文献番号”
（7桁）　（4桁）　（3桁）

さらに、同一文献・資料の中で当該のデータがどの調査スタンドのデータであるのかを識別するための番号（以下調査番号という）を付し、これについても一つの文献・資料に含まれる調査スタンド数を勘案して3桁の数字とした。これにより調査スタンドを特定するための識別コードは次のとおり全部で10桁の数字で表すこととなる。

“調査スタンドを特定する識別コード”＝“調査年代”＋“文献番号”＋“調査番号”
（10桁）　（4桁）　（3桁）　（3桁）

なお、同一の文献・資料の中で調査年代の異なる調査スタンドが混在する場合には、時系列上での植生変化や植物の分布の変化を取り扱う上での便宜を考慮して、調査年代ごとにそれぞれ異なる文献・資料として取り扱うこととした。

③フィルドデータファイルのフォーマットの検討

データファイルのフォーマットを検討するにあたって、登録すべきアイテムと各アイテムに含まれる選択枝あるいは要素（以下カテゴリーという）の内容と数または範囲を決める必要がある。それらを決めたうえで、実際にデータ入力する際のアイテムの順序や桁数（バイト数）を設計した。

なお、本研究の場合には端末から直接にデータ入力する形をとったため、1レコードの長さは72バイト（8ビット）以下に設定しており、使用機種の制約から漢字やひらがなの入力はできないシステムとなっている。

まず、データファイルに登録すべきアイテムを検討した結果、必要なアイテムが基本的に網羅されており他の調査方法によるフィルドデータのアイテムもほぼ包含していることから方形区データのアイテムを採用するのが妥当と判断し、これを基本としてアイテムの選定を行った。その結果20種類のアイテムを登録すべき情報とした。さらに各アイテムについてデータ形式を検討・決定し、カテゴリーの数を勘案して必要バイト数を決めた。選定・決定した各アイテムの内容は次のとおりである。

a．文献番号

調査あるいは文献を識別するためのコードである。文献番号は歴年ごとに、3桁までの、調査あるいは文献ごとに異なる任意の数字を用いることとし、調査年との組み合わせで調査あるいは文献を特定できるようにする。データ形式は数値（実数）型、バイト数は3とする。

b．調査番号

同一の調査あるいは文献の中で、調査スタンドを識別するコードである。調査番号は調査あるいは文献ごとに、3桁までの、調査スタンドごとに異なる任意の数字を用いることとし、調査年、文献番号との組み合わせで調査スタンドを特定できるようにする。データ形式は数値（実数）型、バイト数は3とする。

c．調査年

フィルド調査年を示す。調査年は西暦を用いることとし、文献からの引用の場合にも極力実際のフィルド調査が行われた年を登録する。データ形式は数値（実数）型、バイト数は4とする。なお、調査年が不明の場合には文献・資料の発行年を、調査年が数年にまたがっていて各調査スタンドの調査年が不明の場合には最も新しい年代を調査年とする。

d．調査月

フィルド調査月を示す。データ形式は数値（実数）型、バイト数は2とする。

e．調査日

フィルド調査日を示す。データ形式は数値（実数）型、バイト数は2とする。

f．調査地点名

フィルド調査の地点名を示すコメント欄である。通常“地点名、市町村名”で表示する。データ形式は文字型、バイト数は12とする。

g．調査地点位置

フィルド調査地点の所属する、国土地理院発行の1：25,000地形図を示す。分布図を作成したり特定の範囲のフロラを作成したりする場合に必要となるデータである。調査地点位置は、別に作成する地形図コードにより4桁の数字で登録する。データ形式は数値（実数）型、バイト数は4とする。

h．調査者名

フィルド調査の調査者名あるいは文献の著者名を示すコメント欄である。データ形式は文字型、バイト数は12とする。

i．調査面積（基本的に方形区・帯状区調査の場合に使用）

方形区や帯状区調査の調査面積を示すコメント欄である。単位はm^2、データ形式は文字型、バイト数は4とする。

j．調査方法

調査方法を示す。カテゴリーは「方形区調査」、「帯状区調査」、「植物目録」、「標本」の4種類とし、それぞれを「ＱＵＡ」,「ＢＬＴ」,「ＬＳＴ」,「ＳＰＭ」の記号で識別する。データ形式は文字型、バイト数は3とする。

k．傾斜方向（基本的に方形区・帯状区調査の場合に使用）

フィルド調査地点の傾斜方向を示すコメント欄である。傾斜方向の表現は「Ｎ３０Ｅ」、「Ｓ７０Ｗ」、「Ｓ　５Ｅ」、「　　Ｅ」などのようにする。データ形式は文字型、バイト数は4とする。

l．傾斜角度（基本的に方形区・帯状区調査の場合に使用）

フィルド調査地点の傾斜角度を示すコメント欄である。傾斜角度は2桁の数字で示す。データ形式は文字型、バイト数は2とする。

m．水深

フィルド調査地点が池沼などの場合にその水深を示すコメント欄である。単位はcm、データ形式は文字型、バイト数は4とする。

n．海抜

フィルド調査地点の海抜高度を示すコメント欄である。海抜高度は4桁以下の数字で示す。単位はm、データ形式は文字型、バイト数は4とする。

o．群落名

調査スタンドの群落名を示すコメント欄である。データ形式は文字型、バイト数は28とする。

p．摘要

調査スタンドに関する摘要を示すコメント欄である。文献の頁数や図表番号などを記入する。データ形式は文字型、バイト数は28とする。

q．階層の高さおよび植被率（基本的に方形区・帯状区調査の場合に使用）

調査スタンドの階層ごとの最大の高さと植被率を示すコメント欄である。階層は高木層、亜高木層、低木層、草本第1層、草本第2層（草本層）、コケ層の6層に区分する。高さの単位はm、植被率の単位は%とするが、高木層、亜高木層、低木層の各階層の高さは小数点以下1位で、草本第1層、草本第

2層（草本層）、コケ層の高さは小数点以下2位で表示する。データ形式は文字型、バイト数は各階層の高さがそれぞれ4、植被率がそれぞれ3とする。

r．植物名の表示方法の区分

発表された資料において、植物種が学名で表示されているのか和名で表示されているのかを識別して示す。学名が表示されている場合には「ＡＣＮ」，和名のみが表示されている場合には「ＪＰＮ」の記号で識別する。データ形式は文字型、バイト数は3とする。

s．発表の形態の区分

資料の発表形態を示す。公表されている印刷物の場合には「ＰＢ」、手書きの原稿およびその複写の場合には「ＨＷ」、公表されていない場合には「ＩＤ」の記号で識別する。データ形式は文字型、バイト数は2とする。

t．出現植物名、所属階層、優占度、群度

調査スタンドにおける出現植物名とその所属階層・優占度・群度を示す。分布データと標本データの場合には出現植物名だけの登録となる。

植物名は別に検討・作成される植物分類ファイルのコードを用いて、6桁の数字で示す。データ形式は数値（実数）型、バイト数は一階層一種類の植物につき6とする。

所属階層は高木層、亜高木層、低木層、草本第1層、草本第2層（草本層）、コケ層ごとにそれぞれ「Ｔ1」、「Ｔ2」、「　S」、「Ｈ1」、「Ｈ2」、「　M」の記号で識別する。データ形式は文字型、バイト数は一階層一種類の植物につき2とする。

優占度は「5」、「4」、「3」、「2」、「1」、「＋」、「Ｒ」、「Ｄ」、「Ａ」、「Ｆ」の１０種類の数字または記号で識別する。なお、「Ｒ」は稀、「Ｄ」は優占、「Ａ」は多い、「Ｆ」はしばしば出現を意味する。データ形式は文字型、バイト数は一階層一種類の植物につき1とする。データ形式はデータ処理の中で数値型に変換する。

群度は「5」、「4」、「3」、「2」、「1」の数字で識別する。データ形式は文字型、バイト数は一階層一種類の植物につき1とする。データ形式はデータ処理の中で数値型に変換する。

以上のアイテムの内容をもとに、図2に示すフィルドデータファイルのフォーマット（入力用）を設計した。

（次報へ続く）

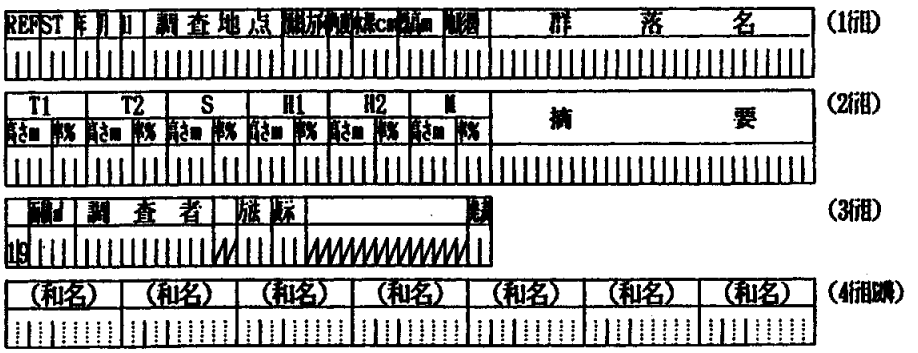

図2　フィルドデータファイルのフォーマット（入力用）