教養としての情報学 序章

玉井哲雄

1　はじめに

東京大学では 1993 年から「情報処理」という授業を，文系理系を問わずすべての学部 1 年生の必修科目とした．その科目の内容は世間から「コンピュータ・リテラシー」を教えるものと見られている．それが科目の狙いの一部であることは確かだが，情報という対象を扱うのに固有な基本的原理や方法に目を向けさせるという大きな目的もある．大体，コンピュータ・リテラシーという言葉が何を指すもののかはっきりせず，あまり好きではないということは別に書いた [2].

さて，2003 年 4 月から高校の普通科目として「情報」が必修になる．その情報の教育を受けた高校生が，2006 年 4 月から大学に入ってくる．それに伴って「情報処理」の内容を変えなければならない．というより大学入学者がある程度は「コンピュータ・リテラシー」をすでに身に着けているものと仮定すれば，大学教養課程における情報教育の内容を一新できる．現在それに向けて，学内グループで議論を進めているところである．

文系理系と共通で必修の授業を行うことを前提として，まず教科書を作ってみたいと考えた．そこで次のようなメモを書いてみた．これは筆者のまったく個人的な構想である．

タイトルは仮に「教養としての情報学」とする．もっとよいタイトルを考えたいが，今のところ思いつかないので取りあえずこうしておく．

1. 趣旨

   - リテラシー教育ではない.
   - 文理共通に使える.
   - 基本的な考え方を伝えたい.
   - 最先端技術 (はやりの技術) も積極的に取り上げるが，それを通して基本概念，技術を教える．たとえば XML を取り上げるとすれば，XML についての解説を書くのではなく，これを題材として，データの蓄積と検索の問題，文書の交換，表示，標準化の問題などを一般的に論じる．すなわち，きちんとした枠組みの中に，流行の技術を位置づける.
   - 情報の面白さと奥の深さを伝えたい.
   - 「体系的」な本ではなく，面白く読めるものとする．たとえば例題を工夫して，ありきたりでない豊かなものとする.
   - 座学的部分と実習部分とを組み合わせる.
   - コンピュータを単なる道具と見るな，というメッセージを伝える.
   - ページ数を多くする.
   - 先行する駒場の参考例:
     - 「知の技法」シリーズ → 一般にも売れる

- Universe of English → 「花子の米国旅行」でもなく Shakespeare でもないことを狙いとしている.
  本屋に平積みされているマニュアル的な解説本でもなく Knuth でもない，というように読み替えればよいか.
- 野矢茂樹: 「論理トレーニング」 → 面白い演習

2. 読者対象: 駒場の 1・2 年生であるが，さらに一般に
   - 「情報」について考えてみたいが，マニュアルやその類いの機械的な文章には辟易している人.
   - 論理的な指向は好きだが，あまりの形式化は敬遠したい人.
3. 内容: 以下は単なる思いつきの羅列

(a) 「伝える」
communicate, 通信，伝達
メディア

(b) 「表現する」
書く，描く，記述，図示
言語，記号，符号，図式
修辞，文章作法

(c) 「考える」
推論，類推，演繹，帰納
論理

(d) 「決める」
選ぶ，判断，意思決定，評価
価値

(e) 「計算する」
数える，演算
計算可能性，計算の複雑さ

(f) 「測る」
計測，計量
統計

(g) 「探す」
探索

(h) 「解く」
求解
方程式，制約，束縛

(i) 「発見する」
パターン，法則

(j) 「理解する」
認識，認知
文章，意味，音声，図形

(k) 「変換する」
換える，翻訳，置き換え
コンパイラ

(l) 「編集する」
編む
テキスト，図，Web ページ，本

(m) 「覚える」
思い出す，記憶，蓄積
DB，記憶装置

(n) 「設計する」
design

(o) 「学習する」
知識獲得，知識適用

思いつくまま 15 の項目を並べてみたが，レベルがそろっているとはいえず，重複する部分もある．何より 15 という数は多すぎる．われわれのグループでは，とりあえず「表現」,「伝達」,「計算」,「検索」,「システム」,「社会」という 6 つの分野に絞って検討しようということになっている．

しかし，このままでは先に進まないので，部分的に草稿を書き始めてみることにした．内容的にもまだごく一部であり，練れてもいないが，SEA の会員諸兄姉にご批判をいただければありがたいと思い，未完成な原稿をお目にかけることにした次第である．以下は上の項目でいえば,「表現」に当たるところの一部である．

2 情報の表現

表現とは

「表現」という言葉は，感情表現，芸術表現などのように人が内面に持つ心理的・精神的なものを外面的な対象として表す行為を指す場合と，文字，音声，図などの何らかの記号によって外部的に表されたものを指す場合とがある．ここでは情報の表現を，現象，事象，事実，規則・法則，などを記号として表すこと，あるいは表したもの，としてとらえる．感情表現や芸術表現の手段は，言語という記号的なものだけでなく，表情，身振り，動作，音楽，絵画，彫刻，など多様である．ここでは記号的な表現に絞ることから，感情表現や芸術表現はとりあえずは除いて考える．ただ情報の表現と感情表現や芸術表現は関連があるので，今後それらに触れることもあるかもしれない．

手段を記号に限ると，表現の定義は情報そのものの定義とほぼ一体化してしまう．情報の定義としてたとえば吉田民人の定義を挙げてみよう．

> 「情報とは，最広義には物質 - エネルギーの時間的・空間的および質的・量的なパタン．最狭義には個体的・集合的な人間主体の意思決定を規定する，伝達された単用的・認知的な外シンボル集合＜有意味な記号集合＞.」[1]

かなり難解だが，最広義の方は宇宙線のパターンや DNA といった自然界の「情報」も含めて考え，最狭義の方は人と人との間で伝達 (communicate) される情報を対象としている．後者は要するに有意味な記号集合といっているわけで，いいかえれば情報は表現されて初めて情報になるのだといえよう．したがって表現の形や方法の考察は，情報を考える上でもっとも基本となるものである．

記号と符号

記号としてまず思い浮かぶのは文字である．文字の中でも数字 (アラビア数字) は今や世界共通に使われている．そこでまず電話番号や郵便番号などの数字による情報表現を考えてみよう．

その前に，記号としての文字の性質に注意しておく．漢字のような表意文字には当然意味が結びつけられているが，表音文字のアルファベットや「かな」にも，音という固有のものが結びつけられている．あるいは「?」のような文字にも「はてな」とか「疑問」という概念が文化的，習慣的に結びつけられている．この結びつけは文化によるので，たとえば日本では○は「よい」，×は「悪い」，△はその間という結びつけが小学校以来一般化しているが，欧文の文脈では必ずしもそのように解釈されないようである．なお，?や○や×や△を記号と呼んで，記号を文字の一部とする言い方もあるが，ここでは逆に記号の一部が文字であると考える．

数字にも数という概念が結びついている．しかし，電話番号や郵便番号で使われている数字は，数として扱われているわけではない．このように固有の意味と結びつけずに使われる記号を，ここでは符号 (code) と呼ぶことにする．モールス信号はトンとツーという長さで区別される 2 つの信号を用いるが，このトンとツーは固有の意味を持たないという点で，典型的な符号といえる．郵便番号や電話番号で使われる数字は，別の記号に置き換えてもかまわない．1 の替わりに?を使い，2 の替わり!を使うというように取り決めても，それほど大きな問題はない．その意味では数字を符号として用いているのである．もちろん，もともと記号すべてにそのような性質があり，?が疑問を表したり 1 が数の「1」を表したりすることに必然性はない．記号にはそのような意味で本来，中立性，代替可能性があり，特定の意味と結びつけられるのは社会の慣習によるのである．

符号化

電話番号は符号としての数字を，たとえば 10 桁並べたものとして表現される．このように固定した符号の集合（今の場合は 0 から 9 までの数字の集合）の要素の並び (符号列) で情報を表現することは，基本的な方法である．符号集合の大きさを n，符号列の長さを l とすると，n^l 個の対象を表現できる．このように情報の対象を一定の符号の組み合わせと結びつけることを，符号化 (coding) という．一般に，言葉は音素あるいは文字の並びで作られるという意味では同じである．ただ，言葉は単語というレベルでとらえても，文というレベルでとらえても，その記号列の長さが不定であるところが電話番号とは異なる．

識別

ここで，表現同士が識別されることと，識別された表現が決まった対象を指すという対応関係を持つことが重要である．このような対応関係を表現から対象への写像という．電話番号で言えば，10 桁の数字列から加入電話への写像が定義されているわけである．

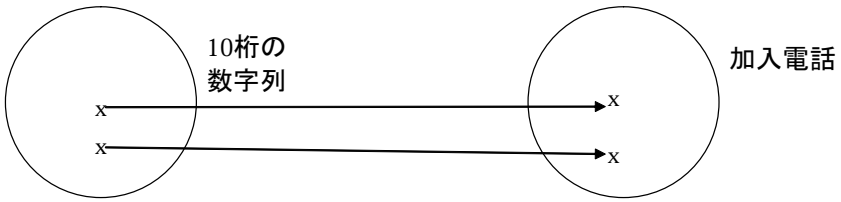

図 1: 表現から対象への写像

写像の基本的な性質として，写像元（定義域と呼ばれることもある）の 1 つの要素（今の場合電話番号の

10 桁の数字列）は，ただ 1 つの写像先（値域と呼ばれることもある）の要素（今の場合は加入電話）に対応している，ということが挙げられる．ソシュール流の言語学で言えば，写像元はシニフィアン（signifiant, 能記とか記号表現と和訳されることもある）に対応し，写像先はシニフィエ（signifié, 所記とか記号内容と和訳されることもある）に対応する．しかし，この両者を二面的に合わせもつものがシーニュ（signe, つまり記号）であるというのだから，われわれのこれまでの記号という言葉の使い方と異なり，ちょっとややこしい．

しかし，ソシュールのような小難しい言語学をもちださず，日常的な言葉の感覚で議論すると，同じ言葉が 2 つ以上の意味を指すことはよくある．これを同音異義語 (homonym) という．一方，違う言葉が同じ意味を指す場合もよくらい，これを同義語 (synonym) という[1]．

写像については同音異義語に対応するものは最初から排除されているが，同義語に相当するものは一般には許される．そのような事例をもたない写像，すなわち写像元の要素が異なれば写像先の要素も異なるものは，とくに単射と呼ばれる．電話番号のような人為的に定められた符号化の体系では，最初から同音異義語や同義語が存在しないように作られる．つまり，記号表現と記号内容の関係は単射な写像となる．

電話番号や郵便番号が数としての性質を持たないことは，それらの間でたとえば

- 四則演算
- 大小関係

が意味を持たないことからも納得できよう．電話番号同士を足したり，2 つの電話番号を比べてどちらが大きいといっても意味がない．

しかし，表現が識別できるためには，2 つの電話番号が同一か否かという判定は意味を持たなければならない．つまり 2 つの番号 a と b に対し，$a = b$ あるいは $a \neq b$ という論理式は意味を持つ．これが「識別できる」ということの基礎である．

少し面倒くさいことを言えば，符号の並びで表現された情報を識別することは，その要素である符号が識別できることに依っている．数字の場合は 10 個の文字の間で $a = b$（または $a \neq b$）の関係が判定できることを前提とする．これは自明なことではない．実際，郵便番号は手書きの数字を機械で読み取って，この判定をしているのである．たとえば  と  は同じと判定するが， とは異なると判定するわけである．

その上で，2 つの記号列 $a_1, a_2, \ldots, a_n$ と $b_1, b_2, \ldots, b_m$ が等しいとは，両者の長さが等しく $(n = m)$，対応するそれぞれの記号が等しい場合 $(a_i = b_i, i = 1, \ldots, n)$ で，その場合に限る．

「近さ」の構造

電話番号や郵便番号に，等しいか等しくないかということ以上の構造はないだろうか．すぐ気づくと思うが，やはり構造はある．それは「近さ」という関係である．電話番号が近いもの同士は，それが指す加入電話の設置場所も互いに近いだろう．

ところで電話番号が近いとは，より正確にはどういう意味だろうか．10 桁の番号の内，9 桁までが一致していたら 8 桁一致しているものより近いだろうか．必ずしもそうではない．桁がより左にある方が近さの判定により大きな影響を及ぼす．そこで 2 つの電話番号の近さの基準として，両者の差の絶対値を取って，それが小さければ近い，大きければ遠いと判断することは考えられる．しかし，電話番号は数ではなく，その間に差とか大小関係は意味がないと言ったのではなかったか．

確かに電話番号の表現に数字の並びを用いたために，10 桁の 10 進数と同じ形となったのは偶然ではあった．0〜9 の代わりに a〜j を使ってもイ〜ヌを使ってもよかった．ただ，個々の記号の間にも何らかの近さ

[1] 同音異義語と対称に書くなら異音同義語そ記すべきだろう．ただし，同音とか異音とかいっても，音のみを念頭においているわけではなく，記号表現として同じか異なるかを問題としている．その意味ではホモニム，シノニムという言い方の方が，対称性もありよいかもしれない．

の性質が存在している必要はある．たとえばaとdの間はaとbの間よりも遠い，というような性質である．これはアルファベットやイロハの順序で対応させることができよう．ただ，aとdの間がaとbとの間より3倍遠いかどうかは判らない．同様にある桁の1単位の違いは，その右の桁の1単位の違いより大きな違いであることは確かであるとしても，その差が10倍かというと，そうとはいえないだろう．

差や絶対値が定義されているのは数（自然数）の世界である．だから，電話番号の差の絶対値が近さを表すというとき，厳密に言えば次のような操作を行っていることになる．

1. 2つの10桁の数字の並びaとbを自然数の世界に写像する．写像をgと書き，aとbのgによる写像先をそれぞれ$g(a), g(b)$と書く．
2. 写像先の数の間の差の絶対値$|g(a) - b(b)|$でaとbとの距離を定義する．
3. 10桁の数字の並びから加入電話への写像をfと書く．$f(a)$と$f(b)$の設置されている場所の近さとaとbの距離が対応しているものとする．

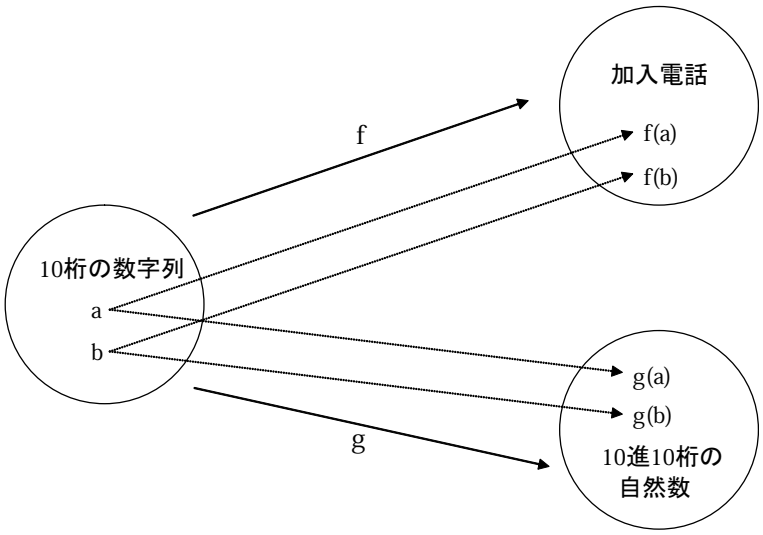

図 2: 2つの写像

階層構造

ここでちょっと待てという声が，聞こえてきそうである．電話番号を10桁の数字の並びとしているが，実は局番と局内番号に分かれているのではないか．郵便番号も，最初の3桁と後の4桁は扱いが違うのではないか．

確かに電話番号の内の局番は加入電話が設置されている地域という地理的な情報を反映しているかもしれないが，下4桁にはそのような対応はなさそうである．上位2桁は，01が北海道から秋田・岩手まで，02が東北の山形・宮城から西は長野まで，03は東京都区内とその周辺，04は都内の残りと，千葉，神奈川，埼玉，という具合に，日本列島の北から南西に地域を分割して決められているらしい．ただ，市外局番一覧表というものを見ると，その桁数は03という2桁のものから，01242のような5桁のものまでまちまちである．そして市外局番の次に，市内局番がある．

とにかく，10桁の数字の並び自身の中に，段階的な構造があるわけである．階層構造とは大きな概念を分割して次のレベルの概念を導き，それをさらに分割して次のレベルの概念を導くという段階的な構造を

いう．電話番号の場合は第 1 段階が市外局番，第 2 段階が市内局番，第 3 段階が局内番号という階層構造を持っている．この構造を明示するために，段階を表す部分列の間にハイフンや括弧を入れるという表記もよく用いられる． 電話番号では，この第 1 段階と第 2 段階の桁数がまちまちだが，これは歴史的な経緯によるものだろう．ただし，市外局番，市内局番というような地理的構造に対応する階層構造があるのは，通常の加入電話の場合である．携帯電話の場合は，仮にこのような階層構造があったとしても地域に対応するものではないだろうし，利用者はその構造を意識していない．間にハイフンなどを入れて区切った表示も，単に記憶の便宜に過ぎないかもしれない．

郵便番号では，第 1 段階が 3 桁，第 2 段階が 4 桁で，その区分は一律である．

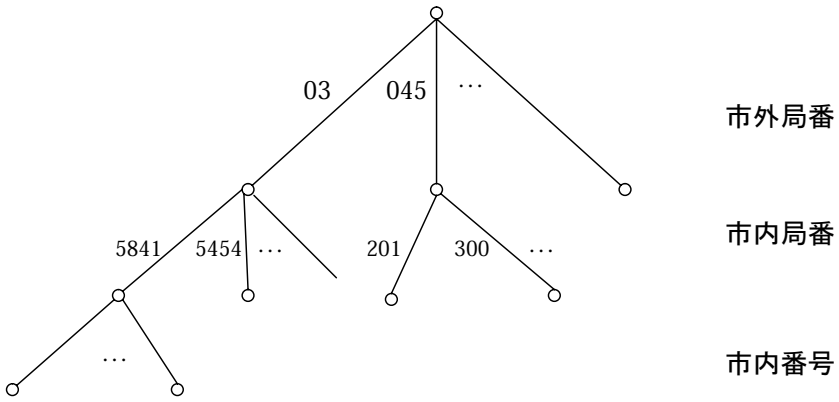

図 3: 電話番号の階層構造

このように部分列が段階を表していて，それを順に配置するという構造は珍しいものではない．たとえば年月日を表すのに，20010911 のような表現をとることがある．最初の 4 桁が年，次の 2 桁が月，次の 2 桁が日を表す．これもやはり左から右に段々小さな単位を置く階層構造になっている．この構造を明示するために，2001/09/11 のように区切りにスラッシュをやピリオドを置く記法もよく用いられる．ただし，ISO の標準は “YYYY-MM-DD” という形式である．つまり年 (4 桁，ただし下 2 桁による省略記法も可)，月 (必ず 2 桁)，日 (必ず 2 桁) と並べ，区切りはハイフンである．このハイフンは省略してよいことになっている．

階層構造の上位レベルを左に，下位レベルを右に置くことは，必然ではない．実際，ヨーロッパでは逆に日月年の順，つまり 11/09/2001 のように書く．面白いことに米国では日本流でもヨーロッパ流でもなく，月日年の順に並べる．つまり 09112001 となる．年月日によってどの流儀で表されているかユニークに決まるものと，そうでなくあいまいなものとがある．とくに米国流とヨーロッパ流の混同が問題だが，たとえば 1213 なら，これは米国式表記が使われていて 12 月 13 日を指していることは明らかである (年の 4 桁は共通なので省略)．しかし 1208 では米国式で 12 月 8 日なのか，ヨーロッパ式で 8 月 12 日なのかあいまいである．

なお，年を下 2 桁で表す表記もよく用いられる．これが 2000 年問題を引き起こしたことは，記憶に新しい．

日本流が大きな構造から小さな構造の順に表現するのに対し，欧米流では異なる順に並べる他の例に，住所表記がある．日本では，都道府県，市町村，丁目，番，号のように並べる．欧米だと，番地，通り，市，州などのように並べる．日本人は構造的な思考が弱いという指摘がよくなされるが，このように大きなものから小さなものへというトップダウンの記述は，階層構造を自然に反映してある意味では合理的だといえよう．

問題１　年月日の米国流とヨーロッパ流の表記で，どちらの表記によるものかがユニークに定まる月日とそうでないものとをすべて示せ．

問題２　図書分類表を調べ，それがどのような階層構造を表現しているか明らかにせよ．

数としての構造

これまで扱った数字列の表現では，識別と近さという構造を考えた．つまり等号と比較という，ごく基本的な演算が定義された集合としての記号列を対象とした．ここではもう少し複雑な数としての構造を考えよう．

すでに挙げた例の中で，年月日は部分的に数としての性質をもっていた．そもそも，われわれが慣れ親しんでいる数の 10 進表記が，情報の表現法の 1 種に他ならない．考えて見れば，数字の並びの位置で 10 進の桁を表すというのは，実に偉大な発明であった．つまり

$$a_1 a_2 \cdots a_n$$

という表現 (a_i は 0〜9) は数として

$$a_1 \times 10^{n-1} + a_2 \times 10^{n-2} + \cdots + a_n$$

を表しているのである．この表現の有用性はたとえば 568 と五百六十八や DLXVIII(ローマ数字) とを比べてみれば，明らかだろう．

年月日の表現の 20010911 は全体として 10 進数を表していないが，最初の 4 桁，次の 2 桁，最後の 2 桁は，それぞれ 10 進数を表している．ただし，月を表す 2 桁は 1 以上 12 以下，日を表す 2 桁は 1 以上 31 以下という制約がある．厳密に言えば，日の上限は月および年の値によって 31, 30, 29, 28 のいずれかに限定される．年の 4 桁は一般には 1 以上 9999 以下だが，日常的には 2000 ± 50 ぐらいの範囲で用いる．

これを月の 2 桁は 12 進数を表していて，12 を越えると年の最下位に桁上がり起こるとみてもよい．同様に日の 2 桁は 31 進数（または 30,29,28 進数）で，上限を超えると月の桁に桁上がりが起こる．

時分秒の表示も同じ構造をしている．たとえば 10 時 48 分 25 秒を 114825 と表現する．これは通常のデジタル時計の表示であり，区切り記号を入れるとすればコロン (:) である．面白いことに，この順序は米国でもヨーロッパでも変わらない．またこれは，コロンを省略できることも含めて，ISO の標準でもある．

最初の 2 桁は 24 進，次の 2 桁は 60 進，最後の 2 桁も 60 進である．ここでは年月日表示の日の桁のようなややこしさはない．この 6 桁をひとまとまりの数の表示とみなすこともできる．

まず左から 6 桁目と 4 桁目は 10 進で桁上がりが起こり，5 桁目と 3 桁目は 6 進で桁上がりが起こる．1 桁目と 2 桁目は 24 進数を 10 進表示しているので，2 桁目が 10 進で，1 桁目が 3 進というような言い方は正確ではない．60 が 10 の倍数なのに対し，24 は 10 の倍数ではないからである．しいて言えば，1 桁目は 3 進でよいが，2 桁目は 1 桁目が 0,1 の時 10 進，2 の時 5 進となる．

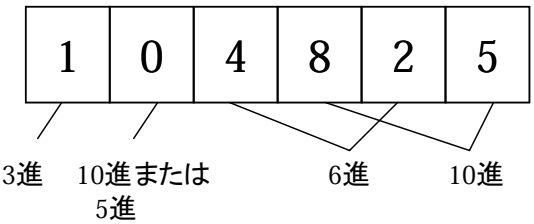

図 4: 時分秒の各桁の表現

このような桁上がり演算のもとで，この時分秒の値に対し加減算が定義できる．たとえば

$$104825 - 085336 = 015449$$

これは 10 時 48 分 25 秒から 8 時 53 分 36 秒を引くと，1 時間 54 分 49 秒になるという演算を表している．左から 3 桁目や 5 桁目の引き算は 6 進のため，上位桁から 6 を借りてきている．

ところで今の説明は，このような桁ごとの桁上がりや桁下がりの演算方法を定義したので時分秒のデータに加減算が定義できたような言い方だったが，これは正しくない．話は逆で，時分秒単位の時刻という値の間に加減算が定義できるが，それをこのような 0-9 の数字による 6 桁表現をした時に，その表現上で個々の桁ごとの桁上がり・桁下がりを含む演算に帰着できるということである．われわれは 10 進表現に慣れてしまっているので，小学校でやる 10 進の足し算や引き算の方法が，足し算や引き算の定義そのものと思い込んでしまうことがあるが，あれもたまたま 0-9 という数字を用いた 10 進表現を使う時に有効な方法に他ならないのである．

ところで，“104825-085336 = 015449” は「10 時 48 分 25 秒から 8 時 53 分 36 秒を引くと，1 時間 54 分 49 秒になるという演算を表している」と言った．ここで微妙な違いに気づいた読者もあるだろう．左辺の 104825 や 085336 は「時刻」を表しているのに対し，右辺の 015449 は同じ表現をしていながら「時間」という別の種類の情報を指している．時分秒の間で加減算ができると言ったが，正確に言えば，次のような演算が可能なのである．

時刻 $-$ 時刻 $=$ 時間
時刻 $\pm$ 時間 $=$ 時刻
時間 $\pm$ 時間 $=$ 時間

ここで引き算 $a - b$ に関して，b が a より大きい時はどうなるだろうか．まずうるさいことを言うと，a と b は通常の数でないから,「b が a より大きい」という意味を定義しておかなければならない．a の時間部分 (最初の 2 桁) を a_h, 分部分 (次の 2 桁) を a_m, 秒部分 (最後の 2 桁) を a_s と書くことにする．b についても同様．その時，次の規則で b が a より大きい ($b > a$) と定める．

1. $b_h > a_h$ なら $b > a$
2. $b_h = a_h$ の時，$b_m > a_m$ なら $b > a$
3. $b_h = a_h, b_m = a_m$ の時，$b_s > a_s$ なら $b > a$
4. それ以外はすべて $b > a$ でない．

ここで a_h, a_m, a_s などは数として扱っている．このように個々に大小関係のあるものの並びに関して，全体の大小関係を部分の大小関係を左から順位適用して決めるというやり方は，広く行われている．それに基づいて昇順（小さい方から大きい方へ）あるいは降順 (大きい方から小さい方へ) に並べたものを，辞書的順序という．

さてずいぶん手間をかけて $b > a$ を定義したが，実は a と b の時分秒表現をそのまま 10 進数とみなしてその数の大小関係で $b > a$ を決めたものと，結果的には同じである．つまらないことに精力を使ったと感じるだろうか，それとも厳密な議論をして気持ちがよいと思うだろうか．

回り道をしたが，$b > a$ の場合の $a - b$ の話である．たとえば今日の日の出が 054230, 日の入りが 175545 だとしよう．この差

$$054230 - 175545$$

を負の数を導入して -121315 と表してもよい．これを時刻 x に足せば，12 時間 13 分 15 秒前の時刻になる．つまり日の出は日の入りの 12 時間 13 分 15 秒前だったと読むのである．しかし，同じ時間を負の数は使わ

ずに，114645 と表してもよい．時刻 x の 12 時間 13 分 15 秒前は，x の 11 時間 48 分 45 秒後と等しいのである．そんな馬鹿なと思うかもしれないが，ここで扱っている時刻の範囲では，日の違いは無視される．要するに日（や月や年や曜日）の表示のない時計と同じである．このような計算を法 (modulo) 計算という．

空間の表現

時間の次は空間である．1 直線上の点を座標という数値で表すのも，一種の符号化といえるかもしれない．しかし，直線上の点は実数に対応している．たとえば線分の左端を 0，右端を 1 とし，その線分上の任意の点に対応する実数 $a(0 \leq a \leq 1)$ を考えた時，a の 10 進表記は一般に桁数が無限になる．そこで桁数を有限に固定して，0.31415 のように表現することが一般に行われる．この表現は，直線上の 1 点の表現というより，1×10^{-5} の幅を持った区間を表していると見ることができる．

一般に表現は有限の記号の有限な組合せで表されるから，本質的に有限である．もちろん記号で「∞」とか「無限」とか書いて無限を表すことはできる．数学の世界では，無限を有限の記号で表すためにさらに精緻な工夫を凝らしてきた．しかしここでは，有限桁の 0.31415 のような表現が空間の点 (の集合) の 1 つの符号化になっていることを指摘するのに留めておこう．この直線上の点を座標で表す方式は，デカルトによって，さらに 2 次元や 3 次元空間の点を表すのに拡張された．平面上の点を表すのに，直交する 2 つの直線，x と y を取り，そこへの射影をとることによって，2 つの数の組を用いるものである．図 5 の点 P は (2.6, 1.8) という数の対で表される．

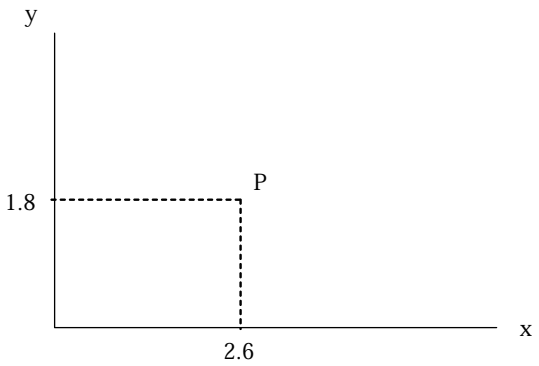

図 5: デカルト座標

このような表現は日常生活でもよく使われる．たとえば囲碁は，19 × 19 の格子の点に石を置いていくゲームである．その格子点は 4 五のように表される．横方向を左から右にアラビア数字で 1 から 19，縦方向を上から下に漢数字で一から十九として表す習慣である．ここで左右とか上下というのは，先手黒番側から盤を見た際の向きをいう．

9 × 9 盤を用いる将棋も同様であるが，面白いことに横方向は囲碁と逆で，右から左に 1 から 9 のアラビア数字をふる．縦は上から下に漢数字を当てることは同じである (図 6 参照)．新聞の将棋欄を見れば,「先手 7 六歩」というような表現で指し手を示しているのが判るだろう．

このように囲碁や将棋やチェスなどの盤面 (board) ゲームは，2 人で交互に指し手を繰り返すことで進行し，その手は盤上のマスないし格子点を表す座標と駒の種類で表現できる．そこで，これを計算機処理に向いたデータ表現にした場合も，データ量はきわめて少なくてすむ．たとえばコンピュータ碁で国際的に使われている SGF(Smart Go Format) では,

```
B[pd];W[cq];B[pq];W[po];B[dd];W[oq];B[or];W[op];B[nq]
```

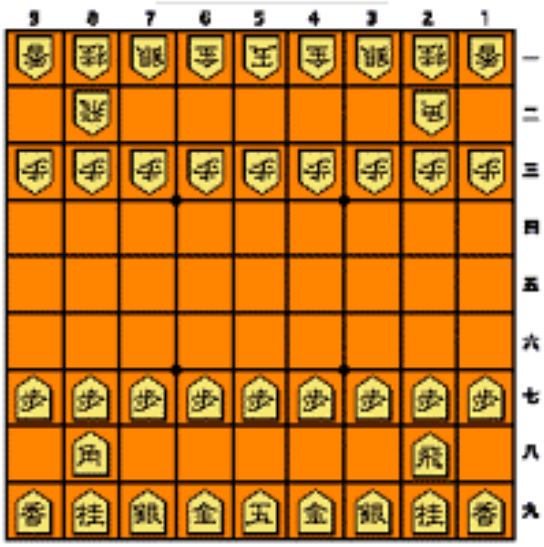

図 6: 将棋の棋譜

のように手の進行を表す．B と W はそれぞれ黒（先手），白（後手）を表し，[pd] などが格子点を表す．横座標も縦座標も a から s までのアルファベットで表し，横方向は左から右，縦方向は上から下に割り当てる．区切りにはセミコロンを用いる．SFG では他に，対局日，対局者名，手についてのコメント，勝敗の結果，など他のデータの記述方法が規定されているが，手の進行という棋譜にとっての本質的な部分は，このようにきわめて簡潔である．

将棋にもさまざまなコンピュータ用の棋譜形式がある．その 1 つ CSA という標準形式では，指し手は “+2726FU” のように表される．ここで＋は先手を意味し，後手なら－である．27 は座標を示し，横軸が 2 縦軸が 7，ただし横軸は右から左に進むことは従来の将棋の棋譜の慣習に従う．したがって 27 は，新聞の棋譜などでは 2 七と表されるものと同じである．次の 26 も同じく 2 六という位置を示す．ここでは駒の動きをあいまいさなく表現しており，先の 27 は指し手前，後の 26 は指し手後の位置を表す．最後の FU は「歩」という駒の種類を表す．だからこの表現は,「先手 2 六歩」に対応する．他に対局者，対局日などのさまざまな情報を記述でき，またある局面の盤面全体を表す表現もある．

このようなコンピュータ向きの表現は簡潔であるが，人間にとっての読みやすさは考慮されていない．しかしこのような形式で記述された棋譜データから，人間にとって判りやすい画像形式の表示を生成することは比較的簡単である．しかもコンピュータ上で表示する場合は，印刷物への表示と異なり，マウスをクリックすることで 1 手ずつ進行させたり，必要なら元に戻したり，任意の手数の局面に飛んだりすることなどが容易に実現できる．

名前

ここでしばらく数字から離れて，一般の文字による表現を考えてみよう．アルファベットやかな漢字による符号化の代表例は，人名，地名，商品名などの名前である．しかし名前に使われる文字は，電話番号における数字のように意味的に中立な記号として用いられるわけではなく，そのためこれを符号 (コード) 化とは呼ばないことが多い．しかし記号列の集合からそれが指し示すもの（人，地域，商品など）の集合への写像として名前が機能するという構造は同じであり，形式的には同じように扱うことができる．

符号化と呼びにくいもう 1 つの理由は，人の名前には同姓同名という現象があることである．地名や商品名でも同じことが言える．つまり記号表現と記号内容の関係は単射な写像とは限らない．実は筆者には，同姓同名 (漢字表記でもかな表記でも) で年齢も同じ，しかも同じ年に同じ大学に入学した人がいる．しかも現在，筆者と同じように大学勤務である．幸いにも職場と専門分野が違うので，さほど頻繁に混乱が起こるわけではないが，それでも過去に何度か混同から来るおかしな事件に遭遇している．

日本の名前は姓と名からなり，多くの国でもこのパターンが多い．姓は例外的な場合を除き，新たに作られることはない．むしろ，結婚に際し夫婦どちらかの姓を選んで新しい家族の姓とし，子供がそれを受け継ぐという現在の日本の制度を続ける限り，希少な姓は確率的に消滅していき，姓の数は単調に減るはずである．一方，名の方は子供が生まれるたびに新たに生成される．だから一般に，姓の種類の数の方が，名の種類の数より多そうである．

ところが日本の姓の数はなんと 29 万種類もあって，世界でも珍しいらしい．これは明治 3 年 (1870) 年に明治政府が平民に苗字を許し，さらに明治 8 年には必ず苗字をつけるように強制し，自分の苗字がはっきりしない場合は新しいものをつけてよいとしたため，爆発的に増えたもののようである．

日本で婚姻に際し改姓することになったのも明治の中頃からで，それまでは女性は結婚後も実家の姓を名乗っていたという．儒教の伝統がまだ生きている中国や韓国では，今でも女性は生家の姓を名乗る．それなのに中国の姓の数は少なく，3000 という説がある．韓国の姓の数はさらに少なく，1985 年の国勢調査で 225 という結果だったという．しかも，金，李，朴の 3 姓で人口の 46%を占めるのだそうだ．日本のように爆発的に姓を創り出すといういい加減なことをしてこなかったからだろう．

日本の名前の付け方も，自由というかいい加減である．出生届の際に名前として使える漢字は，常用漢字だけでなく人名漢字というものが用意されていて量が増やされている（2003 年現在で人名漢字は 285 文字）．その組み合わせ方も自由である．ただ，常用漢字と人名漢字合わせて 2230 文字と，カタカナ・ひらがなは使えるが，アルファベットやアラビア数字は許されない．また読み仮名を振ることになっているが，これと漢字との対応についても何の制限もない．ただ，読み仮名は戸籍には載らず，住民票にのみ登録されるようだ．

保険会社の明治生命が毎年，その年に生まれた赤ん坊に付けられた名前の人気ベスト 10 というのを発表しているが，それによると 2002 年の男の子は上位 10 は「駿，拓海，翔，蓮，翔太，颯太，健太，大輝，大樹，優」，女の子は「美咲，葵，七海，美羽，莉子，美優，萌，美月，愛，優花，凛」（同点のため 11 個）だという．これらのいくつかはどう読んだらいいか判らない．実際，同じ漢字表記にいくつかの読みがあるらしい．

あまりに自由なので，親は子供の命名に際し迷う．そこで字画などの姓名判断に頼り，わざわざ制約条件をつけ選ぶ範囲を狭めて命名するという，面白い現象がある．実際，本屋でその種の棚を見ると，字画による命名という類の本がおびただしく出ているの驚く．

ところがキリスト教，ユダヤ教，イスラム教の文化圏では，子供の名前は基本的に聖人の名前という限られた集合の中から選ぶことが多いようだ．たとえばフランスでは誕生日に結びつけられた聖人の名前から選ぶという習慣が強く，少し以前までは生まれた子供に対し定められた 500 ほどの聖人などの名前の中から名付けることが義務づけられていたという．筆者はある時，米国の学会が主催する会議のプログラム委員会に委員として出席して驚いたことがある．委員は全部で 12 名で筆者以外はすべて米国人．内 2 名が女性だったので米国人男性は 9 名いたことになるが，そのうち 3 人が David だった．アメリカ人は互いに姓でなく名で呼びあうので，紛らわしいことこの上ない．

内包と外延

つまり名前の付け方に 2 つの流儀があることになる．1 つは日本流（と代表して呼んでしまう) で，名前が満たすべき制約条件を決めて，その条件を満たす範囲内で自由に生成する．もう 1 つはフランス流（とこれ

も勝手に代表させてしまう) で，定まったセットの中から選ぶ．付けることのできる名前の集合というものを想定したとき，日本流はその集合に入る要素がもつべき性質を定義するという意味で内包的 (intentional) 定義を与えていることになり，フランス流はその要素を具体的に並べあげるという意味で外延的 (extentional) 定義を与えていることになる．論理学では昔から，概念に対してその「内包」と「外延」という言葉を使う．たとえば「ジャイアンツの選手」というのが内包であり,「清原，上原，高橋 (由) ,...」などと並べあげるのが外延である．

命名規則

付けられる名前の集合についての内包的定義の記述は，命名規則 (naming rule) と呼ぶことができる．その典型は競走馬の命名規則である．日本の競走馬は，カタカナで 9 文字以内と定められている．さらにすでに登録されている名前と同一のものはもちろん，登録をすでに抹消されているものでも抹消後 5 年を経過していないものは付けられないそうだ．さらに有名な馬名（ＧＩ優勝馬や国際的に保護された馬名等）や馬名として不適当またはふさわしくないものも付けられないというあたりになると，規則としてはややあいまいになる．

生物には学名というものが付けられる．これはもちろん競走馬のように個体に付けられる名前ではなく，種に付けられるものである．学名の基となるはリンネ式の階層分類である．すべての生物が分類の上位から下位に向けて，界，門，綱，目，科，属，種という段階で分類される．たとえば人類は動物界脊椎動物門哺乳綱サル目ヒト上科ヒト属ホモ・サピエンス種である．学名はこの分類に基づいたラテン語表記で，まず属名を名詞で (頭文字は大文字)，次に種小名を形容詞またはその相当語 (小文字）で記述する．これが命名規則である．人名が姓と名からなるように，ここでも二名法が取られている．

プログラミング言語の識別子

コンピュータ・プログラムを記述するためのプログラミング言語は，典型的な人工言語である．プログラムを書く際には，プログラムや変数や関数に名前をつけなければならない．それらの名前を総称して識別子 (identifier) という．名前だから識別が重要なのは当然だが，とくに相手がコンピュータだから，ホモニムでも前後の文脈で判断するという融通が利かない．

たとえばC というプログラミング言語の識別子の命名規則は，次のようになっている．

> 1 文字以上の任意の長さの文字列．文字列に許される文字は，先頭は大文字か小文字のアルファベット，2 文字目以降は大文字か小文字のアルファベットか数字か_（下線）．

競走馬の名前のように長さ制限がないので，意味をよく表すように工夫した名前が付けられる．識別子にかなや漢字の使用を許すプログラミング言語もある．

しかし，古いプログラミング言語では長さの制限があった．たとえば今でも技術計算の分野ではよく使われる Fortran では，6 文字以内と規定されていた．しかも使える文字はアルファベットの大文字と数字だけで，小文字は使えなかった．Fortran77 までその制限があったが，Fortran90 と呼ばれる言語仕様では小文字の使用が認められ，長さの上限も 31 文字に広げられた．

長さの制限がなかったりゆるかったりすれば，名前がぶつからないようにするのは少し楽にはなるが，大きなプログラムを多人数で開発する際に，すべての識別子に重複が起こらないように管理するのは大変である．プログラマもそのようなことに余計な神経を使わないで，プログラミング作業を行いたいはずである．そこでほとんどのプログラミング言語で採用されている巧みな仕組みとして，名前の有効範囲 (scope) という概念がある．

プログラムは何らかの単位から構成される．この単位はプログラミング言語によって呼び方も大きさもまちまちであるが，ここではそれをモジュールと呼ぶことにしよう．変数をあるモジュールで宣言した時，その名前はそのモジュールの中だけで有効で，別のモジュールに同じ名前の変数があっても，それは別のものを指す，というのが名前の有効範囲の考え方である．モジュールの中にモジュールがあるという階層構造を許す場合が多いが，その時も，あるレベルのモジュールで宣言された名前の有効範囲は，そのモジュール自身かそれに含まれるモジュールの中（それがまたモジュールを含んでいればそれも含む) とする．これはある意味で，名前を文脈に応じて判断していることになり，あいまいさをなくしながらも，すべての名前を文字列として区別しなければならないという制約をはずすうまい工夫である．

構造的な表現

これまで扱ってきた電話番号のような符号列も人名も生物の学名も，対象となるものを指すための記号表現だった．それらをひっくるめて，改めて「名前」と呼んでもよい．しかし，ものを指す方法は名前によるばかりではない．そのものが持つ性質を並べ挙げて同定するというやり方もある．たとえば「イチロー」という名前である人を指す代わりに,「マリナーズという野球チームのレギュラー選手で，守備位置はライトで，打順は 1 番」というような具合である．

このように対象を定める性質の項目を属性という．属性には種別と値がある．イチローの例で言えば,「所属チーム」という属性種別の値が「マリナーズ」であり,「守備位置」という属性種別の値が「ライト」であり,「打順」という属性種別の値が「1 番」である．このように属性の種別とその値の対の集まりという構造で情報を表現することは，きわめてよく行われる．

たとえば日頃，申込書の類を書かされることが多い．それらには名前，住所，電話番号，性別，年齢，メールアドレスなどの欄がある．欄は属性の種類に対応し，そこに書き込むものが属性の値である．最近は申込書という紙に記入する代わりに，インターネット上の Web のページから申し込む場合のように，指定された欄にキーボードから文字を打ち込んだり，メニューリストからマウスでクリックして属性の値を選ぶことも多くなった．

名前で指すとの別の方法として，属性と値の列挙があるとこの節の話を始めた．それなのに申込書の例では名前も属性となっているのはどういう訳だろうか．実際，名前で 1 つに定まらない場合，さらに社員番号とか学生証番号といった識別符号を属性に入れることもしばしばある．つまり，このように属性で対象を表現するのは，それを名前で指す代わりというわけでは必ずしもなく，むしろその対象のもつさまざまな情報をまとめて表現する手段と考えたほうがよい．

****

これから先，申込書のようなデータを集めたデータベースの話，それと等価な表現としての n 組や表の話，属性の値がまた構造を持った表現であるような階層構造，その表現法の 1 つとしての XML の話などを考えているが，長くなるばかりなので，SEAMAIL の原稿としてはこの辺でやめにする．

参考文献

[1] 吉田民人. **情報と自己組織性の理論**. 東京大学出版会, 1990.

[2] 玉井哲雄. 国際的情報社会に立ち向かう. 浅野攝郎，他, 編, **東京大学は変わる–教養教育のチャレンジ**, pp. 100–115. 東京大学出版, 2000.