JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

OLEG ARKHIPOFF

Taxonomie et sémantique : un projet de recherches

Journal de la société statistique de Paris, tome 117 (1976), p. 230-245

<http://www.numdam.org/item?id=JSFS_1976__117__230_0>

TAXONOMIE ET SÉMANTIQUE : UN PROJET DE RECHERCHES

Première partie : Étude formelle des nomenclatures

Oleg ARKHIPOFF

Administrateur de l'I.N.S.E.E.

Dans cette première partie, les nomenclatures sont essentiellement envisagées sous l'angle d'un calcul formel. Après avoir précisé une terminologie et constaté que la famille des nomenclatures définies sur un même ensemble (nomenclatures homobases) constitue un treillis, on donne des algorithmes automatiques permettant de déterminer la nomenclature la moins fine (respectivement la plus fine) de toutes celles qui sont plus fines (respectivement moins fines) que deux nomenclatures données. Dans le premier cas, on obtient la nomenclature la plus économique pour une enquête à objectifs multiples et, dans le second, on établit le niveau le plus fin permettant de comparer des enquêtes différentes. Passant aux nomenclatures hétérobases, on étudie la notion de M-*injection, cas particulier d'une application multivoque, d'un ensemble nomenclaturé dans un autre.*

In first part, nomenclatures are mainly viewed as a formal calculation. After defining a terminology and noting that the family of nomenclatures defined on one set (homobase nomenclatures) forms a lattice, automatic algorithms are given, allowing to single out the least fine nomenclatures (the finest respectively) among the ones which are finer (less fine respectively) than two given nomenclatures.

In the first case, we get the most economical nomenclature for multipurpose surveys and, in the second case, we establish the finest level allowing to compare different surveys. As for heterobase nomenclatures, we study the M-*injection notion, a particular case of a multivocal application, from a set with nomenclature to another one.*

In diesem ersten Teil sind die Nomenklaturen im Wesentlichen unter dem Gesichtspunkt des formellen Rechnens betrachtet. Nachdem man eine Terminologie festgelegt hat und bei dieser Gelegenheit feststellte, dass die Gruppe der Terminologien für eine gleiche Einheit ein « lattice » bildet, gibt man automatische Algorhytmen, die ermöglichen die Nomenklatur zu bestimmen, die am wenigsten detailliert ist (beziehungsweise die am meisten detailliert ist) von allen, die mehr detailliert sind, (beziehungsweise weniger detailliert sind) als zwei gegebene Nomenklaturen. Im ersten Falle erhält man die Nomenklatur, die am rationnellsten ist für eine Untersuchung mit vielen Objektiven und im zweiten Fall schafft man die Basis, die am raffiniertesten ist und die gestattet die verschiedenen Untersuchungen zu vergleichen. Wenn man auf die Nomenklaturen übergeht, die für verschiedene Gruppen bestimmt wurden, so kann man die Möglichkeit einer Uebertragung von einer Gruppe auf eine andere untersuchen.

Journal de la Société de statistique de Paris, tome 117, n° 3, 1976.

SOMMAIRE

Le présent essai, dont le but est essentiellement heuristique et dont les conclusions restent provisoires, est issu de préoccupations différentes quoiqu'en fin de compte convergentes. Les mots « statistique » ou « économie » n'y apparaissent pas, mais on conviendra facilement que nomenclature et signification jouent ici, comme ailleurs, un rôle primordial.

La première préoccupation est celle du comptable national œuvrant en Afrique : nous avons ainsi été conduits à analyser la nomenclature des biens et services, dite « Courcier » [1] : l'idée directrice de cette systématique est, croyons-nous, la succession bien connue : primaire — secondaire — tertiaire — , combinée avec le (vague) préordre que constituent les biens et services en tant qu'inputs les uns des autres ; il fallait ensuite traduire la nomenclature douanière, dite de Bruxelles, bien plus fine que la Courcier, en cette dernière, d'où quelques problèmes d'articulation et d'interprétation. Vers 1969, nous avons retrouvé le même problème, en plus complexe, lorsqu'il s'est agi de confectionner une nomenclature de biens et services articulée à la fois sur la nomenclature de Bruxelles (la seule que connaissaient les industriels enquêtés), et sur la C. I. T. I. pour tenir compte des recommandations de l'O. N. U. ; en effet cette nomenclature devait faire l'objet d'une proposition pour une nomenclature standard des biens à annexer au Plan comptable O. C. A. M. [2] et servir à une enquête particulière auprès des industriels dans le but d'établir un tableau d'échanges inter-industriels [3] ; naturellement le tout devait s'intégrer sans peine dans le cadre de la comptabilité nationale.

De ces quelques travaux nous sont nées diverses réflexions que nous avions rassemblées dans un brouillon d'article, avec pour thème l'idée que la théorie du domaine à nomenclaturer *dictait* la nomenclature à servir et que la confection de celle-ci ne pouvait être livrée à l'amateurisme et aux joies littéraires, comme cela arrive quelquefois, du moins si nous nous reportons à notre expérience personnelle. Ajoutons que nous nous sommes aussi largement convaincus que c'était la tâche particulièrement ingrate et sans lustre. C'est de cette première ébauche, intitulée « la confection rationnelle des nomenclatures », que nous sommes partis pour rédiger la première partie de la présente étude.

Notre première approche était essentiellement formelle et nous soulignions qu'une théorie sécrétait généralement une équivalence *sui generis* dont il convenait de tenir impérativement compte. Mais, depuis longtemps déjà nous étions également intrigués par l'énigme de la signification ; et nous pensions qu'une nomenclature, avant tout, signifiait

quelque chose mais cela d'une manière qui nous semblait bien obscure. C'est un coin du voile de ce mystère que nous avons tenté de soulever dans la seconde partie de cette publication.

Un troisième facteur, accessoire peut-être, mais déterminant, a été à l'origine de la seconde partie de cette étude, sinon de toute l'étude elle-même : l'étude du concept de bien-être national débouche directement sur la question de savoir comment déduire d'une manière *acceptable* un ordre des préférences collectives à partir d'un ensemble d'ordres individuels de ces mêmes préférences; nous sommes alors arrivés à la conclusion qu'il y avait impossibilité systématique, dans tous les cas de figure, d'agréger convenablement un ensemble donné de relations d'ordre individuelles, voire de relations individuelles quelconques, tant soit peu complexes. Or, la notion d'ordre est primordiale et fondamentale : c'est une des premières notions acquise par l'enfant (notion d' « avant-après », puis ordre vicariant, puis transitivité [4]) et on rencontre constamment cette notion dans toutes les branches de la connaissance. Aussi nous nous sommes légitimement inquiétés de ce qu'il en était en logique, pour le passage des vérités individuelles à la vérité tout court [5], nous réservant pour plus tard d'examiner le problème de l'agrégation des significations.

Ce dernier problème n'a évidemment pas échappé à Ferdinand de Saussure, mais les trois (si nous comptons bien) « réponses » qu'il donne nous ont laissés plutôt perplexes. Parlant de ce bien collectif qu'est le fameux Trésor de la langue, de Saussure écrit : « Si nous pouvions embrasser la somme des images verbales emmagasinées chez tous les individus, nous toucherions le lien social qui constitue la langue. » Et, ailleurs : « la langue existe dans la collectivité sous la forme d'une somme d'empreintes déposées dans chaque cerveau, à peu près comme un dictionnaire dont tous les exemplaires, identiques, seraient répartis entre les individus ». Quant aux mécanismes mêmes de cette agrégation, il se borne à dire : « les associations ratifiées par le consentement collectif, et dont l'ensemble constitue la langue, sont des réalités qui ont leur siège dans le cerveau » [6].

Aussi, sans prétendre résoudre le problème du sens, ni répondre à une question qui a mis en échec le grand linguiste genevois, nous avons rédigé la seconde partie de cette étude d'un point de vue sémantique, en essayant de définir la signification qui s'attache à une nomenclature et d'analyser les données qui président à l'agrégation du sens.

Ajoutons ceci : une mauvaise lecture de Saussure pourrait faire croire qu'un linguiste (en fait un sémanticien) *ne doit pas* faire intervenir les nomenclatures dans ses théories. Disons qu'il n'en est rien, comme nous tâcherons de le montrer le moment venu.

En résumé, donc, la présente étude se subsdivise en deux parties assez indépendantes l'une de l'autre. La première, plutôt formelle, est consacrée à l'analyse des nomenclatures et propose une certaine terminologie. La seconde aborde le problème du sens en s'attachant surtout à ce qu'on pourrait appeler la signification lexicale vue sous l'angle d'une analyse par nomenclature; cette dernière partie ne prétend à aucun moment produire des conclusions péremptoires et définitives.

La démonstration des théorèmes est donnée en annexe (1) et les références bibliographiques, signalées par des numéros entre crochets à la fin de cette étude. En ce qui concerne le vocabulaire utilisé, nous ne chercherons pas à distinguer entre systématique, taxonomie ou taxinomie, etc.; nous confondrons également, le plus souvent, langue et langage et nous ne chercherons pas non plus à opposer sème à sémème par exemple.

1. Laquelle sera publiée, avec la seconde partie de cette étude, dans le prochain numéro de ce Journal.

Première partie

ÉTUDE FORMELLE DES NOMENCLATURES

1. *Introduction*

Nous nous proposons, dans cette première partie, d'examiner d'un point de vue essentiellement pratique — sans malice en quelque sorte — les principales propriétés de l'ensemble des nomenclatures définies sur un même ensemble E (nomenclatures homobases).

Étant donc donné un ensemble E, que nous appelerons *base*, une *nomenclature* M de base E sera tout simplement une partition de cet ensemble. En d'autres termes, M se présente comme une famille de sous-ensembles M_i non vides de E, sous-ensembles dits *rubriques* de M, tels que l'intersection de deux quelconques d'entre eux soit toujours vide et dont la réunion est la base E elle-même.

Soit $\mathcal{N}(E)$ l'ensemble des nomenclatures *(homobases)* définies sur E. On notera I la nomenclature dont toute rubrique est constituée par un seul élément de E, et O, celle à une seule rubrique, E lui-même. I sera dit la nomenclature *fondamentale* et O la nomenclature *triviale*.

Il existe une autre définition de la nomenclature considérée cette fois-ci comme une suite de partitions successivement emboîtées les unes dans les autrs [7], et effectivement c'est bien comme cela qu'apparaît en pratique toute nomenclature concrète [8].

Le nombre des rubriques de M est souvent dit *niveau* de M. Deux nomenclatures sont égales si elles ont les mêmes rubriques et il est clair qu'un même niveau correspond généralement à plusieurs nomenclatures distinctes. On notera enfin que la notion de niveau ne prend un sens concret que lorsque le niveau est fini.

Rappelons enfin, et ce point est capital, que se donner une nomenclature sur E, c'est-à-dire une partition de E, *équivaut* à se donner une relation d'équivalence sur cet ensemble E.

2. *La notion d'articulation*

Se pose fréquemment le problème suivant : étant données deux nomenclatures M et N (homobases), retrouver les rubriques de l'une en combinant des rubriques de l'autre. Si ce problème est soluble (sans une tierce nomenclature), on dit que ces deux nomenclatures sont bien articulées l'une sur l'autre.

On dira donc que M est *articulée* sur N — et l'on écrira MaN — *si toute rubrique* M_i *(ou* N_j*) est nomenclaturée par des rubriques de* N *(ou* M*) ou bien est incluse dans une rubrique de* N *(ou* M*).*

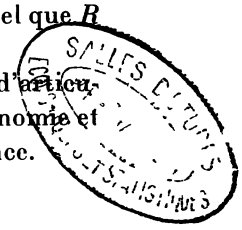
Théorème 01 : MaN équivaut à la condition :

$$(\forall i)\,(\forall j)\,(M_i \cap N_j = \varnothing \quad ou \quad M_i \cap N_j = M_i \quad ou \quad M_i \cap N_j = N_j).$$

L'articulation est réflexive et symétrique. Enfin, toute nomenclature s'articule sur la nomenclature fondamentale et la nomenclature trivale.

Rappelons que $(\forall x)R$ signifie : R est vrai pour tout x, et $(\exists x)R$: il existe un x tel que R soit vrai; on notera *ou* la disjonction et, plus loin, & la conjonction logiques.

Ce premier théorème donne donc deux définitions équivalentes de la notion d'articulation qui se trouve être ainsi précisée. C'est là un terme constamment utilisé en taxonomie et on peut noter encore que ce n'est ni une relation d'ordre, ni une relation d'équivalence.



3. *Produit de deux nomenclatures*

Étant données deux nomenclatures quelconques, M et N, on constate en général que celles-ci ne sont pas articulées. Comment produire une troisième nomenclature P telle que MaP et NaP?

C'est là un problème important, car il peut se concevoir au minimum comme un problème d'agrégation de nomenclatures et même comme une agrégation de significations lexicales.

Une solution à ce problème est donnée par la notion de *produit* (ou d'intersection) de deux nomenclatures : $P = M \times N$, que nous définirons comme suit :

$$P = \{M_i \cap N_j;\ M_i \in M \ \&\ N_j \in N \ \&\ M_i \cap N_j \neq \varnothing\}$$

la notation $\{x;\ R\}$ signifiant « l'ensemble des x vérifiant la proposition R ».

Nous définirons encore $M \geqslant N$ par $M \times N = M$ et nous dirons que M est *plus fine* que N, ou encore que M est une *sous-nomenclature de* N. Énonçons :

Théorème 02 : Le produit de deux nomenclatures est encore une nomenclature; il est idempotent, commutatif et associatif. En outre, $M \times I = I$ et $M \times O = M$. La relation $\geqslant$, dite *finesse* est une relation d'ordre et la borne supérieure de tout ensemble de deux nomenclatures est le produit de celles-ci.

Ce théorème donne les propriétés essentielles du produit : quand on l'effectue, on n'a pas à se soucier de l'ordre dans lequel on considère les nomenclatures, ni du nombre de celles-ci. La relation de finesse est bien une relation d'ordre, comme le suggérait sa définition et la nomenclature fondamentale apparaît comme le plus grand et la nomenclature triviale comme le plus petit élément de $\mathcal{N}(E)$; pour toute nomenclature M, donc :

$$I \geqslant M \geqslant O.$$

Dire que $P = M \times N$ est la borne supérieure de M et de N signifie que, premièrement: $P \geqslant M$ et $P \geqslant N$ et que, deuxièmement, quelle que soit Q, $Q \geqslant M$ et $Q \geqslant N$ implique $Q \geqslant P$. On peut encore considérer que le produit apporte non seulement une réponse à un problème d'articulation mais encore et surtout une solution à un problème d'agrégation. Le mode d'agrégation consiste, ici, à prendre la borne supérieure de l'ensemble des nomenclatures à agréger; ce mode d'agrégation n'est évidemment pas arrowien mais répond au principe d'unanimité : si $M = N$, alors $P = M \times M = M$, ce qui n'est autre que la propriété d'*idempotence* du produit [5].

Nous allons donner dans un instant un algorithme permettant de calculer par voie informatique le produit de plusieurs nomenclatures. Mais il convient de noter que l'usage de l'algorithme présuppose la connaissance explicite des rubriques de M et de N et cela au moyen de la nomenclature fondamentale I *supposée connue.*

Il est clair que la connaissance de I détermine $\mathcal{N}(E)$ et que c'est là une condition *sine qua non.* Mais il est non moins sûr que la détermination de I fait *toujours* problème et — ajouterions-nous — cette détermination est, en règle générale, toujours provisoire (l'étude de la nomenclature fondamentale est ce que B. H. Dussart, à la suite d'Aymonin, appelle la *taxonomie,* par opposition à la *taxinomie* « science des arrangements, des lois de la classification » et la *systématique,* étude de la « filiation des organismes vivants ou ayant vécu » — *in* « Bienfaits et méfaits de la systématique en écologie », bulletin de la Société zoologique de France. Rappelons qu'ici nous utilisons ces termes de façon indifférenciée).

Dans la pratique, il arrive souvent que l'on ne « connaisse » pas I mais qu'on dispose d'un ensemble de nomenclatures non articulées — d'un corpus en quelque sorte — censées être homobases : le problème de la détermination de I n'est plus alors, à proprement parler, un problème de nomenclature, *mais bien un problème d'agrégation sémantique* (on peut ici opposer l'approche normale du statisticien qui est de définir la nomenclature par extension à l'approche de l'économiste qui travaille surtout en compréhension).

Avant de donner l'algorithme du produit, soulignons encore une dernière propriété intéressante : la définition du produit peut encore se mettre sous la forme $x(M \times N)y$ équivaut à xMy & xNy, pour tous x et y, l'écriture xMy, par exemple, signifiant x est équivalent à y modulo M, c'est-à-dire que x et y appartiennent à une seule et même rubrique de la nomenclature M, laquelle, rappelons-le encore une fois peut être considérée aussi comme une relation d'équivalence.

Intuitivement, cette propriété veut dire que si l'on considère une théorie de E qui est une conjonction logique d'une théorie à laquelle est associée la nomenclature M et d'une autre à laquelle est associée une nomenclature N, il sera naturel d'associer $M \times N$ à la théorie composée.

Le produit de nomenclatures est aisé à mettre en œuvre. Mais quand le nombre des nomenclatures est élevé et leur encombrement certain, il est intéressant de disposer d'un algorithme permettant de conduire les calculs à bonne fin.

La figure 1 donne l'ordinogramme de l'algorithme et le tableau 1, le programme correspondant écrit en FORTRAN.

La donnée de E revient à énumérer ses éléments (NN) et celle des nomenclatures M et N à préciser pour tout élément de E, *outre son numéro* NN, le numéro NI de la rubrique M_i qui le comprend et le numéro NJ de la rubrique N_j qui lui correspond : on se donne ainsi un jeu de cartes, chacune correspondant à un élément de E et donnant NN, NI et NJ pour l'élément considéré.

Un jeu d'essai de 32 cartes (puissance de E) : $NN = 1, 2, \ldots, 32$, première colonne — effectue le produit d'une nomenclature M à 16 rubriques, $NI = 40$ à 55, deuxième colonne, avec une nomenclature N à 8 rubriques, $JN = 70$ à 73 et 75 à 78, troisième colonne. Le résultat P est donné par la quatrième colonne : $K = 300$ à 325. Les numérotations NN, NI, NJ ayant été fait au hasard, il se trouve que P est presque identique à I.

4. *La relation de finesse*

Le problème de l'articulation n'a évidemment pas toujours une solution unique et le produit ne constitue qu'une d'entre ces solutions, mais c'est certainement une solution élégante, parce que facile à mettre en œuvre. Le produit présente encore un autre avantage, celui d'introduire de façon naturelle la relation de finesse entre nomenclatures.

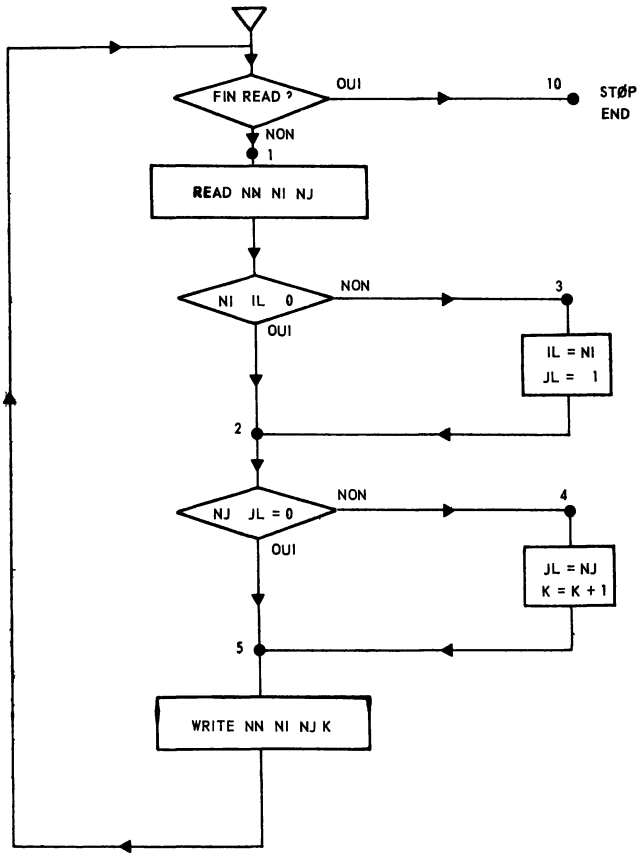
Examinons brièvement les propriétés de cette relation. Mais, auparavant, rappelons deux définitions [9] :

$$M > N =_d (\forall_i)\ (\exists_j)\ (M_i \subset N_j),$$
$$M \vdash N =_d (\forall_j)\ (\exists_i)\ (M_i \subset N_j).$$

La première de ces deux relations (lesquelles sont définies pour des familles d'ensembles *quelconques*) est dite *finesse au sens intérieur* et la *seconde finesse au sens extérieur*. Il est facile de montrer que ce sont là deux ordres non totaux.

On peut maintenant énoncer.

Figure 1 - ANALYSE DU PRODUIT DE DEUX NOMENCLATURES NI ET NJ



INITIALISATIONS

IL = - 1

K = 299

K est le numero de la rubrique de P = M N , P sera donc numerotee a partir de 300

Il est inutile d initialiser JL, cela etant fait au passage de la premiere carte (NN = 1, NI = 40, NJ = 75 dans l'exemple choisi) On a pris pour IL et JL des nombres negatifs puisque jamais, une nomenclature n'a de numero negatif

Tableau 1 - PROGRAMME DONNANT LE PRODUIT DE DEUX NOMENCLATURES AVEC JEU D ESSAI

```
FORTRAN IV G LEVEL  18                  MAIN

0001                    INTEGER * 2 IL,JL
0002                    WRITE(6,20)
0003                 20 FORMAT(1H1,'RESULTATS',///)
0004                    IL=-1
0005                    K=299
0006                  1 READ(5,40) NN,NI,NJ
0007                 40 FORMAT(3I2)
0008                    IF(NI-IL)3,2,3
0009                  3 IL=NI
0010                    JL=-1
0011                  2 IF(NJ-JL)4,5,4
0012                  4 JL=NJ
0013                    K=K+1
0014                  5 WRITE(6,50) NN,NI,NJ,K
0015                 50 FORMAT(4I4)
0016                    IF(NN-32)1,10,10
0017                 10 STOP
0018                    END
```

RESULTATS

1	40	75	300
2	40	75	300
3	40	78	301
4	41	70	302
5	41	70	302
6	42	71	303
7	42	75	304
8	42	76	305
9	42	76	305
10	43	71	306
11	44	70	307
12	44	70	307
13	44	72	308
14	45	72	309
15	45	75	310
16	46	70	311
17	46	71	312
18	46	76	313
19	47	72	314
20	47	72	314
21	47	73	315
22	48	70	316
23	49	71	317
24	50	70	318
25	51	70	319
26	52	70	320
27	53	70	321
28	54	76	322
29	54	77	323
30	55	72	324
31	55	72	324
32	55	76	325

Théorème 03 : Dans $\mathcal{N}(E)$, la finesse équivaut à la finesse au sens intérieur et implique la finesse au sens extérieur et l'articulation. Bien plus, dire que M est plus fine que N revient à dire que M est articulée sur N et plus fine au sens extérieur que N. L'ensemble des nomenclatures définies sur E, ainsi ordonnée par la finesse, possède un plus grand élément : I (la nomenclature fondamentale) —, et un plus petit : O (la nomenclature triviale).

Donc, la relation de finesse est une relation particulièrement significative, puisqu'elle implique deux autres relations notables, dont celle d'articulation, et surtout, elle précise la notion d'emboîtement : chez Volle *et alii*, une nomenclature s'entend comme une suite N^1, N^2, ..., N^k de partitions de E emboîtées, c'est-à-dire telles que

$$N^1 \geqslant N^2 \geqslant \ldots \geqslant N^k \text{ [7] [8].}$$

L'emboîtement dont il vient d'être question se justifie surtout par des considérations pratiques : facilité d'accès d'une nomenclature, facilité d'identification. Mais on se rend déjà compte que la mise en relief d'une telle relation d'ordre ne peut se fonder exclusivement sur la seule commodité et que d'autres relations d'ordre sont également possibles, telle la finesse au sens extérieur, telle la finesse en rubrique (M est *plus fine en rubriques* que N si, par définition, le nombre de rubriques de M est supérieur ou égal à celui de N — c'est un préordre qu'implique la finesse), etc. Nous avouerons qu'il est difficile de trouver des justifications plus théoriques et cette même difficulté se retrouve dans toute la problématique sémantique : commodité de description ou mise en évidence de relations fondamentales inhérentes à la nature des choses? (Certes, il y a des emboîtements plus artificiels que d'autres, mais on ne saurait concevoir un emboîtement parfaitement arbitraire et qui soit néanmoins opérationnel.)

On peut seulement noter que la finesse est intimement liée au produit et comme on le verra, à la somme de deux nomenclatures, et que $M \geqslant N$ équivaut à « xMy implique xNy » ce qui, intuitivement signifie qu'une théorie plus forte qu'une autre devra disposer d'une nomenclature plus fine.

Il est encore utile de définir la relation non réflexive et non symétrique suivante : $M \rightarrow N$ si N est obtenu à partir de M en fusionnant deux (et deux seulement) rubriques de M; plus précisément, en numérotant convenablement les rubriques de M et de N,

$N_2 = M_1 \cup M_2$ et $N_i = M_i$ $\qquad (i = 3, 4, \ldots, n)$.

Ceci suppose que, M ayant n rubriques, N n'en a plus que n—1.

On dira encore que M *couvre* N [10] si M est distincte de N et plus fine que N et s'il n'existe aucune nomenclature P, distincte de M et N et telle que $M \geqslant P \geqslant N$. Nous noterons alors cette relation M/N. On a, évidemment, $M \widetilde{\geqslant} N$ dès que $M \rightarrow N$.

Théorème 04 : Dire que M couvre N revient à dire que N est obtenue à partir de M après fusion de deux rubriques de celle-ci et de deux seulement :

$$M \rightarrow N \Leftrightarrow M/N.$$

Donc le passage d'une nomenclature à une autre immédiatement supérieure s'effectue en fusionnant deux rubriques *ceteris paribus*. Ce théorème répond à une ancienne de nos préoccupations (*cf.* deuxième partie, § 3 *in fine*; elle est également utilisée *in* [7].

5. *Somme de deux nomanclatures*

Il est maintenant intéressant d'adopter l'optique inverse de celle qui nous a conduits à la définition du produit de deux nomenclatures et de tenter de construire une nomencla-

ture S plus grossière (mais pas trop) que deux nomenclatures données M et N, et qui resterait articulée sur chacune de celles-ci.

La définition de la somme S de deux nomenclatures est moins parlante que celle du produit P. L'idée est de rendre équivalents deux éléments x et y de E qui sont directement, ou indirectement par le biais d'éléments tiers, équivalents-M ou équivalents-N. Nous poserons alors : x est équivalent-S de y s'il existe une suite d'éléments $z_0, z_1 \ldots, z_{k+1}$ tels que $z_0 = x$ et $z_{k+1} = y$ et z_i équivalent-M ou équivalent-N de z_{i+1}, c'est-à-dire, de manière plus stylisée, xSy équivaut à

$$(\exists z_1)\ (\exists z_2)\ \ldots\ (\exists z_k)\ (xMz_1\ \&\ z_1Nz_2\ \&\ \ldots\ \&\ z_kNy).$$

Une telle définition peut être d'ailleurs étendue à des relations M et N quelconques, mais il conviendra alors de spécifier l'alternance des M et N, par exemple : $MMNMNMM\ldots N$. Énonçons un lemme qui va simplifier tout cela.

Théorème 05 : Si M et N sont réflexives, leur somme l'est aussi. Dire que M est transitive équivaut à dire que $M + M = M$.

Dès lors, si l'on se limite à la somme des nomenclatures, une suite MMM par exemple équivaut à M. D'autre part la réflexivité de la relation d'équivalence permet aisément de passer de, par exemple, $xMy\ \&\ yNz$ à $xNx\ \&\ xMy\ \&\ yNz\ \&\ zMz$: on n'a donc pas lieu de se soucier outre mesure des alternances MN ou NM dans la définition de la somme des nomenclatures.

Théorème 06 : La somme de deux nomenclatures est encore une nomenclature. Cette somme est idempotente, commutative, associative et vérifie la propriété d'absorption

$$M + (M \times N) = M \times (M + N) = M.$$

Ce théorème montre que la somme s'utilise avec les mêmes facilités que le produit quant à l'ordre des facteurs et le nombre de ceux-ci.

La propriété d'absorption, peu parlante en elle-même, a surtout une signification technique : elle permet d'affirmer que $\mathcal{N}(E)$ constitue ce qu'on appelle un treillis (ou encore un lattis ou un réseau ordonné) :

Théorème 07 : La famille $\mathcal{N}(E)$ des nomenclatures homobases définie sur E, munie de la relation de finesse, est un treillis doté d'un plus petit et d'un plus grand élément. La borne supérieure de deux nomenclatures M et N est $M \times N$ et la borne inférieure est $M + N$. En résumé :

$$O \leqslant M \leqslant I,$$
$$M \times N = \vee \{M, N\},$$
$$M + N = \wedge \{M, N\},$$
$$M \geqslant N \Leftrightarrow M + N = N \Leftrightarrow M \times N = M,$$
$$M \times N \geqslant M, N$$
$$M + N \leqslant M, N$$
$$P \geqslant M\ \&\ P \geqslant N \Rightarrow P \geqslant M \times N$$
$$P \leqslant M\ \&\ P \leqslant N \Rightarrow P \leqslant M + N$$

où $\wedge E$, $\vee E$, $\Rightarrow$, $\Leftrightarrow$ signifient respectivement : borne inférieure de l'ensemble E, borne supérieure, implication et équivalence logiques.

Figure 2.–ORDINOGRAMME DE LA SOMME DE DEUX NOMENCLATURES

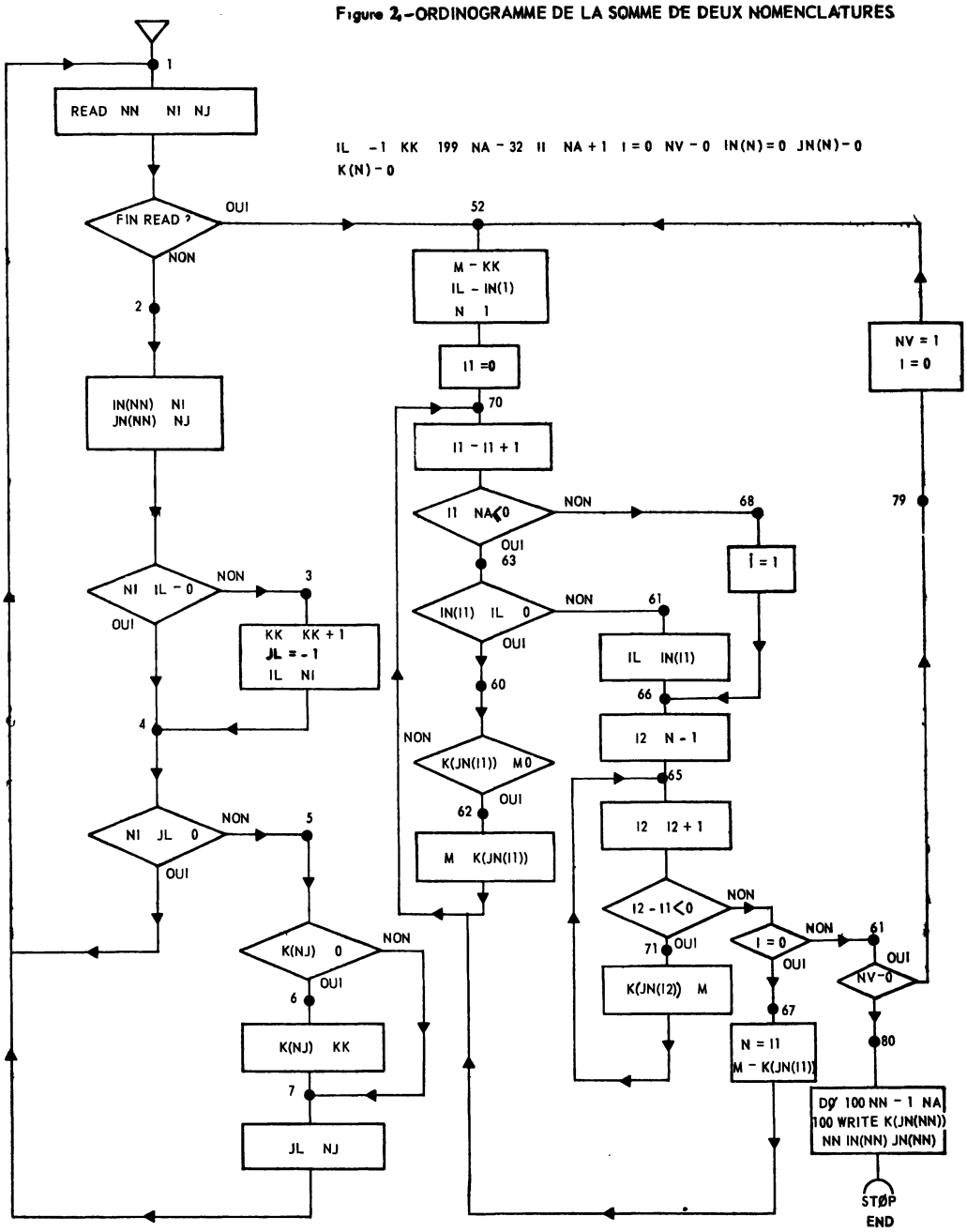


Tableau 2 - PROGRAMME DONNANT LA SOMME DE DEUX NOMENCLATURES

```
FORTRAN IV G LEVEL  18                  MAIN                DATE = 70322

0001                 DIMENSION IN(32),JN(32),K(90)
0002                 IL=-1
0003                 KK=199
0004                 NA=32
0005                 I=0
0006                 II=NA+1
0007                 NV=0
0008                 DO 40 N=1,32
0009                 IN(N)=0
0010              40 JN(N)=0
0011                 DO 41 N=1,90
0012              41 K(N)=0
0013               1 READ(5,50)NN,NI,NJ
0014              50 FORMAT(3I2)
0015                 IF(NN-II)2,52,52
0016               2 IN(NN)=NI
0017                 JN(NN)=NJ
0018                 IF(NI-IL)3,4,3
0019               3 IL=NI
0020                 KK=KK+1
0021                 JL=-1
0022               4 IF(NJ-JL)5,1,5
0023               5 IF(K(NJ))7,6,7
0024               6 K(NJ)=KK
0025               7 JL=NJ
0026                 GO TO 1
0027              52 M=KK
0028                 IL=IN(1)
0029                 N=1
0030                 I1=0
0031              70 I1=I1+1
0032                 IF(I1-NA)63,63,68
0033              63 IF(IN(I1)-IL)61,60,61
0034              60 IF(K(JN(I1))-M)62,70,70
0035              62 M=K(JN(I1))
0036                 GO TO 70
0037              61 IL=IN(I1)
0038              66 I2=N-1
0039              65 I2=I2+1
0040                 IF(I2-I1)71,72,72
0041              71 K(JN(I2))=M
0042                 GO TO 65
0043              72 IF(I)69,67,69
0044              67 N=I1
0045                 M=K(JN(I1))
0046                 GO TO 70
0047              68 I=1
0048                 GO TO 66
0049              69 IF(NV)80,79,80
0050              79 I=0
0051                 NV=1
0052                 GO TO 52
0053              80 DO 100 NN=1,NA
0054             101 FORMAT(4I6)
0055             100 WRITE(6,101) NN,IN(NN),JN(NN),K(JN(NN))
0056                 STOP
0057                 END
```

La démonstration de ce théorème est classique (*cf.* [10], par exemple). Nous devons ajouter, pour être complets que le produit et la somme sont partout définis sur $\mathcal{N}(E)$, ce qui est évident et constitue néanmoins une condition impérative pour la validité du théorème 07. Les treillis possèdent de nombreuses propriétés, dont, en particulier l'inégalité modulaire et les inégalités distributives : nous ne les citons que pour mémoire, car nous n'en ferons jamais usage ici.

La signification de la somme en termes de théories de l'ensemble E est délicate à expliciter. Très grossièrement, on peut dire que la nomenclature somme convient à une théorie qui en généralise deux autres.

La détermination de la somme de deux nomenclatures est une opération qui, dans tous les cas, est pénible. Un algorithme est donc ici indispensable. La figure 2 donne l'ordinogramme de cet algorithme et le tableau 2 le programme correspondant. Les symboles sont les mêmes que dans l'algorithme du produit et, en particulier, les cartes ont même

Tableau 3 - SOMME DE DEUX NOMENCLATURES, JEU D'ESSAI

1	40	77	200
2	40	78	200
3	40	70	200
4	41	72	201
5	41	72	201
6	41	71	201
7	42	75	202
8	42	73	202
9	42	73	202
10	42	75	202
11	43	73	202
12	43	75	202
13	43	73	202
14	44	74	200
15	44	70	200
16	45	69	205
17	46	71	201
18	46	76	201
19	46	71	201
20	47	71	201
21	47	76	201
22	47	76	201
23	48	71	201
24	49	71	201
25	48	76	201
26	49	73	202
27	49	73	202
28	49	73	202
29	49	73	202
30	50	71	201
31	50	71	201
32	50	71	201

dessin. *Il importe ici de souligner que les cartes doivent être impérativement présentées dans l'ordre des numéros NI croissants* (ou NJ, en inversant les symboles : la somme est commutative).

6. *Nomenclatures hétérobases*

On est souvent conduit à l'emploi simultané de nomenclatures définies sur des bases différentes.

On peut naturellement définir le produit cartésien de deux nomenclatures comme le produit cartésien des deux équivalences correspondantes et l'on sait que ce produit est bien une équivalence sur l'espace $E \times E'$, E et E' *étant les deux bases. On sait que toute équivalence sur* $E \times E'$ *peut se mettre sous le forme d'un produit d'équivalences sur* E *et* E' (*cf.* [11] E II, 46, § 8).

Une application pratique du produit cartésien peut se trouver dans les nettoyages de fichiers, où l'on définit pour ce faire des « cases » d'incompatibilité. Mais l'intérêt théorique en est manifestement bien plus grand, comme on le verra dans la seconde partie de cette étude, le point essentiel étant ce qu'on appelle quelquefois un problème de représentation d'une structure globale en structures combinées en forme de produit cartésien.

L'exemple du célèbre tableau de Mendeleiev vient au crédit d'une telle conception des choses. Notons encore le principe de non-lacunarité énoncé par le grand chimiste qui montre ce que peut donner une nomenclature bien conçue.

On peut encore noter qu'une comptabilité ordinaire peut être considérée comme une nomenclature à trois dimensions : nature des comptes, nature des opérations et dichotomie débit/crédit.

Pour terminer cette première partie on se posera maintenant le problème suivant : étant données deux ensembles E et E' la théorie conduit à considérer une application f de E dans E'; une nomenclature s'imposant dans E, quelle nomenclature s'impose dans E', compte tenu de f? La réponse à ce problème ne peut être, à ce stade, très précise et nous nous bornerons aux généralités ci-après.

On se propose d'étudier les applications $f : X \to Y$, qui d'une nomenclature définie sur un ensemble X donnent une image, notée $f(M)$:

$$f(M) = \{f(M_i);\ M_i \in M\}$$

qui soit elle-même une nomenclature sur Y. Ici, une « application » pourra être éventuellement multivoque, c'est-à-dire que $f(x)$ pourra comporter plus d'un élément de Y [9].

Nous dirons que f est une *M-injection* si

$$\text{non } xMy \Rightarrow f(x) \cap f(x') = \varnothing$$

Nous appellerons *noyau injectif* N_f de f, l'ensemble des nomenclatures M de $\mathcal{N}(X)$ pour lesquelles f est M-injective. Nous appellerons encore *équivalence associée* à f notée M^f, l'équivalence traditionnelle

$$xM^fy \Leftrightarrow f(x) = f(y)$$

et qui, manifestement, reste une relation d'équivalence quand f n'est plus univoque mais multivoque.

Énonçons maintenant

Théorème 08 : La condition nécessaire et suffisante pour que f soit une surjection est que $f(M)$ soit un recouvrement.

Théorème 09 : La condition nécessaire et suffisante pour que l'image $f(M)$ soit une nomenclature est que f soit une M-injection surjective.

Il apparaît donc maintenant que le transport d'une nomenclature d'un ensemble sur un autre exige au minimum de la correspondance donnée f, dont il a été question *in limine*, qu'elle soit surjective. Donc le problème général posé au début du paragraphe peut ne pas être soluble (mais il est évident que son énoncé devra aussi être précisé davantage).

Nous établirons maintenant le lien qui existe entre le concept de M-injection avec la notion d'injection [11] et celle, un peu moins classique, d'application semi-univoque [9].

Rappelons qu'une *injection* (univoque) est une application qui à deux éléments distincts fait correspondre deux images distinctes, c'est-à-dire

$$x \neq y \text{ implique } f(x) \neq f(y)$$

et, puisque, $f(x)$ et $f(y)$ se réduisent à un élément de Y chacun, ceci se réécrit comme

$$x \neq y \text{ implique } f(x) \cap f(y) = \varnothing$$

On peut généraliser, à partir de là, l'injection au cas des applications multivoques; et une injection (univoque) est une injection multivoque.

Rappelons encore qu'une *application semi-univoque* est définie par :
$f(x) \cap f(y) \neq \varnothing \Rightarrow f(x) = f(y)$ [9]. Il est clair qu'une injection multivoque (donc une injection) est semiunivoque. On peut maintenant énoncer :

Théorème 10 : Une injection (univoque ou multivoque) surjective est M-injective pour toute nomenclature M. Il s'ensuit automatiquement que $N_f = \mathcal{N}(X)$.

Donc une M-injection apparaît comme une généralisation de l'injection surjective multivoque, elle-même généralisation de l'injection surjective, au même titre que l'application semi-univoque.

On peut reprendre cette classification des applications en termes de noyaux injectifs. Énonçons :

Théorème 11 : Si le noyau injectif d'une application contient M, il contient l'ensemble (M) de toutes les nomenclatures moins fines que M.

Théorème 12 : La condition nécessaire et suffisante pour que f soit une injection multivoque est que son noyau injectif contienne la nomenclature fondamentale de X : I —, c'est-à-dire qu'il se confonde avec $\mathcal{N}(X)$.

Théorème 13 : La condition nécessaire et suffisante pour que f soit une surjection semi-univoque est que son noyau injectif contienne la nomenclature associée.

Résumons : f est une surjection si et seulement si son noyau contient la nomenclature triviale; c'est une M-injection si et seulement si M fait partie de ce noyau; c'est une application semi-univoque surjective si et seulement si la nomenclature associée appartient au noyau; c'est une injection surjective multivoque si et seulement si le noyau comprend la nomenclature fondamentale. Enfin, si f est une injection surjective, le noyau comprend toujours la nomenclature fondamentale (mais la réciproque n'est pas vraie, car il faut établir l'univocité).

Le théorème 11 permet encore d'énoncer

Théorème 14 : Si le noyau injectif d'une application f contient les nomenclatures M ou N, il contient leur somme; s'il contient M et N, il contient leur produit $M \times N$. Si f est surjectif, $f(O) = O$ et si f est injective (multivoque ou univoque), $f(I) = I$.

Enfin, si M et N font partie du noyau de f, alors

$$f(M + N) = f(M) + f(N)$$

et

$$f(M \times N) = f(M) \times f(N).$$

La notion de M-injection est certainement précieuse pour l'analyse lexicale, comme on le verra au chapitre 5 de la seconde partie.

La suite de cet artiele paraîtra dans le prochain numéro du Journal de la Société de statistique de Paris.