La gestion informatique des structures socio-professionnelles et démographiques pendant la révolution industrielle (1800-1850)

par

Claude DESAMA

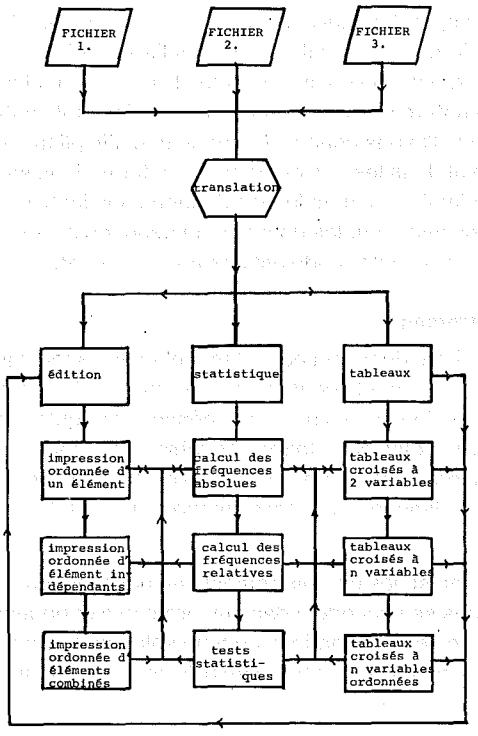
Université de Liège - BELGIQUE

La communication que nous avons ptésentée à l'occasion de ce Congrès International Informatique et Sciences Humaines, expose en termes succincts, les traits généraux du logiciel d'exploitation de listes nominatives de recensement par l'ordinateur : le système LEGIA.

Le champ d'expérimentation et d'application a été constitué de deux listes nominatives de la population de la ville de Verviers (1806 et 1846), siège, pendant la première moitié du 19e siècle d'une Révolution Industrielle «à l'Anglaise».

Le système LEGIA est un ensemble de programmes intégrés qui, au départ d'une structure d'enregistrement définie, analysent le contenu du fichier et en combinent les éléments. LEGIA comprend plusieurs fonctions (calcul, édition, tableaux) qui apparaissent sous la forme de routines ou de sous-routines dans les différents programmes. Conçues de façon modulaire, les composantes de chaque procédure sont, pour la plupart, interchangeables, ce qui confère à l'ensemble une grande souplesse de fonctionnement.

1. LE SYSTEME LEGIA



Pour mieux comprendre, voyons-en l'organigramme :

On y distingue très nettement deux phases : l'une, préparatoire, que nous avons appelée la *translation* dont l'objet est d'organiser la structure d'enregistrement des records en vue de l'exploitation; l'autre, opératoire, qui offre plusieurs options de traitement, combinées ou non entre elles selon les besoins de l'utilisateur.

La combinaison des routines ne se heurte à aucune contrainte à caractère technique ou programmatique. Ainsi, il n'existe aucun obstacle à faire effectuer, dans un même programme, une impression ordonnée d'un élément et d'un tableau croisé à n variables: il suffit de juxtaposer les deux routines en définissant les variables. Toutefois il convient, en cette matière aussi, de respecter un plan de traitement qui tienne compte de la logique interne de l'analyse que l'on veut faire du document et des résultats que l'on désire obtenir : en bref de ce que les chercheurs américains appellent le *Research Design.*

Par ailleurs, il apparaît à l'évidence que certaines routines sont complémentaires et qu'il faut tirer parti de cette complémentarité dans un souci raisonnable de l'optimalisation des traitements.

A. LA TRANSLATION OU L'ORGANISATION DE L'ENREGISTREMENT

Nous appelons *TRANSLATION* l'opération qui consiste à transformer une structure d'enregistrement quelconque en un record dont l'organisation interne permet l'application des programmes standards propres au système *LEGIA.* La translation comprend,selon le cas, un ou plusieurs programmes dont les fonctions varient suivant la nature des transformations à opérer. Elle entraîne au moins une copie du fichier à traiter. Telle que nous l'avons conçue, la translation n'implique aucune modification du contenu de l'enregistrement sauf dans les cas où certaines informations seraient volontairement ignorées dans le traitement. Il s'agit par exemple, des données nominatives figurant dans les fichiers dont l'exploitation porte exclusivement sur les données numériques ou encore d'éléments codés (professions, provenances, ...) dont seul le code est pris en considération.

a/ L'organisation de l'enregistrement

Si l'on excepte la création du critère de tri, qui pose des problèmes spécifiques, la phase de translation vise, en définitive, à mieux organiser le contenu des records. Le but recherché est une plus grande économie de place, préoccupation qui n'a rien d'académique lorsqu'on sait la taille des fichiers que l'on peut constituer avec les protocoles de notaires, les tables des registres paroissiaux ou encore les listes nominatives de recensements des grandes villes. Dans ce domaine, il n'existe aucune contrainte liée à l'operating system ou au langage de programmation : le chercheur choisit l'architecture qui lui convient le mieux.

Pour notre part, nous avons organisé les enregistrements en nous inspirant de deux principes :
premièrement les informations se succèdent dans un ordre fixe et permanent;
deuxièmement il n'existe entre les informations aucune solution de continuité. En fonction de ces deux postulats, la séquence définitive du record se présente comme suit : critère de tri/référen-

ce/identification/données en longueur fixe/données en longueur variable.

Les données en longueur fixe sont regroupées dans *la première partie du record* de façon à connaître immédiatement leur position initiale. Si elles étaient précédées des informations en longueur variable, il nous faudrait recourir, *dans chaque cas,* à certaines fonctions PL/1 (9) pour obtenir ce résultat. Pareille démarche irait, sans nul doute, à l'encontre de la simplification recherchée.

Pour atteindre la forme définitive de l'enregistrement, il est nécessaire d'effectuer deux ou trois opérations selon que le record est composé de zones de longueur fixe et/ou de zones de longueur variable. Ces opérations, dont l'ensemble constitue la translation, sont les suivantes :

a) la création de critères de tri;
b) l'organisation des zones de longueur fixe;
c) l'organisation des zones de longueur variable.

Leur ordre de succession n'est bien entendu pas obligatoire et leur importance respective varie selon la nature et le contenu de l'enregistrement. Il va de soi que la bonne qualité de la translation, et partant son intérêt, suppose que l'on ait une vision déjà élaborée des traitements ultérieurs. Faute de quoi, elle devrait être inévitablement recommencée en cours de travail.

B. LES PROGRAMMES STANDARD

Si l'on ne craignait pas de succomber à la mode terminologique, on qualifierait volontiers de PACKAGE l'ensemble des procédures LEGIA en aval de la translation. Il s'agit, en effet, de programmes standard qui s'appliquent à une structure d'enregistrement donnée où le contenu n'est pas déterminant. Toutefois, par référence à SPSS, à OSIRIS ou à la bibliothèque des programmes économiques et statistiques en provenance de la banque fédérale américaine (FED) (14), force nous est d'admettre que l'universalité d'application de LEGIA n'en est encore qu'à ses débuts. Aussi éviterons-nous une confusion, qui pourrait paraître ambitieuse, en parlant simplement de *programme standard.*

L'organisation des enregistrements et des programmes ne sont sans doute pas indispensables au traitement des listes nominatives de recensement de Verviers (1806-1846). Il est clair cependant que l'investissement en «software» est d'autant mieux rentabilisé qu'il s'applique à d'autres fichiers justiciables d'exploitations analogues. C'est le cas notamment du recensement de Liège en l'an IX. Le coût/travail que représentent la conception, la rédaction et la mise au point de programmes est trop élevé, en effet, pour qu'on se satisfasse d'une utilisation qui ne serait que spécifique et sans lendemain.

Ce souci nous a amené à concevoir non seulement une uniformisation des structures d'enregistrement, – c'est la translation que nous venons de décrire – mais encore des programmes d'exploitation organisés de façon modulaire.

Pourquoi ce surcroît qui peut apparaître comme une complication gratuite ? Pour la raison, aisément compréhensible, que chaque fichier ne réclame pas nécessairement tous les types d'exploitation mais aussi afin de garder aux programmes *LEGIA* une souplesse optimale d'utilisation qui fait généralement défaut aux *PACKAGES* classiques

La raison d'être de *LEGIA* est de fournir aux historiens et aux praticiens de la documentation sociale un instrument susceptible de répondre à des préoccupations dont nous savons la variété. Elle vise donc à être sinon la «panacée informatique» tout au moins un plus grand commun dénominateur capable de réduire, autant que faire se peut, la spécificité des problèmes et des solutions.

1. La fonction d'édition

La fonction d'édition est loin d'être secondaire dans la chaîne des traitements. Elle ne se borne pas à imprimer les données sous une forme de bonne présentation et répond à d'autres impératifs que celui de l'élégance. En fait, il s'agit de regrouper les informations selon des critères qualitatifs, soit pour mettre en évidence tel ou tel caractère, soit pour donneur au chercheur ou au lecteur du fichier un moyen de consultation. A cet égard, la fonction d'édition complète la banque de données et facilite son utilisation pratique. Initialement, la fonction d'édition a été appliquée à un type de fichier et de structure d'enregistrement bien précis. Néanmoins les divers modules de la fonction n'ont pas de vocation spécifique et certains d'entre eux peuvent aisément servir aux éditions d'autres fichiers, *a fortiori* lorsque leur structure d'enregistrement a été soumise à la procédure de translation.

a. L'ordonnance des informations

1. Le rôle du tri

Si les procédures de tri jouent dans les traitements en général un rôle important, celui-ci est décisif dans la fonction d'édition. Il faut savoir en effet que l'ordre selon lequel se déroule une impression est déterminé par le choix des critères de tri dans la procédure de SORT et non dans le programme d'édition proprement dit. Ce dernier organise l'information préalablement triée et met en page l'impression elle-même. C'est au moment du tri que s'opère l'ordonnancement des données et leur combinaison éventuelle.

On distingue généralement deux types de tri : le tri unidimensionnel et le tri hiérarchisé. Comme son qualificatif l'indique, le premier ne concerne qu'une seule variable que l'on trie, soit selon un ordre ascendant, soit selon un ordre descendant (par ex. un tri alphabétique des patronymes de AA à ZZ ou de ZZ à AA). Le second type englobe l'ensemble des tris qui portent sur une pluralité de variables. Il est hiérarchisé dans la mesure où l'on doit obligatoirement fixer la succession des tris, chacun de ceux-ci réalisant une désagrégation supplémentaire des informations.

Le nombre de branches constituant cette arborescence n'est pas, bien entendu, lié à une structure binaire. Le principe du tri hiérarchisé est de désagréger les ensembles formés par le tri précédent et de créer de nouveaux ensembles qui peuvent à leur tour faire l'objet d'une désagrégation. La limite de ce processus est atteinte lorsque l'on a utilisé successivement tous les critères de tri possibles. Il s'agit là, bien entendu, d'une limite théorique car, en pratique, il serait absurde de combiner des variables qui n'ont pas de lien logique entre elles. Opérer un tri selon l'ordre alphabétique des patronymes et le nombre d'enfants mâles de moins de douze ans ne présente en soi aucun intérêt puisqu'il n'existe entre ces deux variables auncune relation de discrimination ou de complémentarité.

En vue des diverses éditions de la liste nominative de recensement nous avons toujours eu recours à des tris hiérarchisés même lorsqu'on ne souhaitait mettre en évidence qu'une seule variable. La raison en est qu'aucun critère de tri n'a assez de pouvoir discriminant, par lui-même, pour lever toutes les ambiguïtés possibles. Si l'on prend la variable qui opère la sélection la plus stricte, le patronyme, on constate, dans des fichiers où la dispersion des noms est aléatoire, que les homonymes sont nombreux et que le recours au premier voire au second prénom est absolument indispensable. L'avantage du tri hiérarchisé sur le tri unidimensionnels réside en ceci qu'il permet de créer des sous-ensembles homogènes où les identités sont connues et acceptées en hypothèse.

Quant au tri unidimensionnel, il est utilisé, en ordre principal lors de traitements intermédiaires (vérification d'enregistrement, listing de travail) ou encore pour l'impression de listes sélectives. Citons, à titre d'exemple, la liste des professions telles qu'elles figurent dans le document avec en regard le code numérique correspondant. Davantage qu'une différence de valeur ou de qualité, c'est donc la fonction assignée à l'édition qui départage le recours au tri unidimensionnel ou au tri hiérarchisé.

2. Le choix des variables.

Etant donné le rôle éminent du tri dans la fonction d'édition et la place qu'il occupe en amont des traitements, le choix des paramètres revêt une grande importance. Autrement dit, il est nécessaire d'établir la liste des variables à mettre en œuvre et leurs combinaisons éventuelles. D'une manière générale cependant celles-ci doivent tenir compte de trois nécessités : la production de listings de travail ou de référence; l'impression de listes sélectives; l'édition du fichier dans son ensemble sous les formes les plus appropriées à une consultation aisée (souci lié à la constitution de la banque de données et à son utilisation).

1/ Listings de travail et/ou de référence

On peut imaginer toutes les variétés possibles de listings de travail et les multiplier à l'envi. Le libre-arbitre du chercheur n'est, en l'occurrence, limité que par les contraintes de coût.

2/ L'impression de listes sélectives

Contrairement aux listings de travail, les listes sélectives dont il est question à présent sont des documents définitifs qui, le cas échéant, peuvent faire l'objet d'une publication.

3/. L'édition du fichier dans son ensemble

Indépendamment des aspects purement documentaires, l'impression d'un fichier complet présente un intérêt analogue à celui des listes sélectives. Comme ces dernières, elle permet d'effectuer des contrôles ou de contribuer à l'enrichissement des informations, mais, cette fois, de manière exhaustive. Elle peut être aussi l'amorce d'études spécifiques d'une nature certes différente mais qui participent du même principe.

2. La fonction calcul et les tableaux

Bien qu'il s'agisse de sous-routines distinctes, les fonctions calcul et l'impression de tableaux sont si étroitement associées dans les faits qu'il nous est apparu plus rationnel de les analyser ensemble

comme nous l'avons fait précédemment pour les modules de tri et d'impression.

a. La fonction calcul

Il n'est pas question de décrire ici les nombreuses fonctions arithmétiques et mathématiques que comporte le PL/1 comme d'ailleurs tout langage évolué. Ces fonctions ne jouent, en effet, qu'un rôle accessoire dans la mesure où elles n'exercent aucune influence sur la conception et l'organisation du système *LEGIA*. Elle n'interviennent, éventuellement, qu'au moment de l'opération de calcul elle-même.

Les modules que nous groupons sous le nom générique de «fonction calcul» se répartissent en deux types de sous-routines : les comptages et les calculs statistiques.

1. Les comptages

Qui dit grand nombre implique presque nécessairement le recours aux méthodes d'analyse quantitative dont la plus élémentaire, – et souvent la plus éclairante –, est le comptage. Entendons par là une addition des occurrences d'une variable ou d'un ensemble de variables combinées. D'emblée nous nous trouvons donc confrontés avec le problème du choix des variables et, par conséquent, des décisions à prendre concernant les critères de tri. Bien que la matière à traiter soit différente, les contraintes sont analogues à celles que nous avons déjà exposées à propos de l'impression des listes (31). Il est toutefois une exigence, propre à toute analyse de données chiffrées, à laquelle nous ne pouvons nous soustraire : la valeur significative des résultats.

Nous ne parlons pas, bien entendu, de l'intérêt ou de l'absurdité sur le plan logique de telle ou telle combinaison de variables mais de leur signification statistique. Il existe en effet des seuils en deçà desquels toute interprétation de données quantitatives n'est plus fondée, ou, à tout le moins. s'effectue avec des précautions autres que de langage. Les abaques mis au point par Liorzou fournissent, à cet égard, des échelles extrêmement précieuses qui évitent l'écueil d'un jugement subjectif (32).

En matière d'analyse quantitative, le choix des variables doit nécessairement prendre en compte ce facteur supplémentaire et prévoir des combinaisons qui restent significatives.

Le principal mérite d'une procédure informatique intégrée comme le système *LEGIA* réside dans le passage d'une démarche déductive fondée sur diverses hypothèses, implicites ou explicites, relatives à la structure de la population verviétoise, à une démarche inductive. Partant de données globales et de quelques regroupements généraux (par ex. les secteurs économiques) l'ordinateur poursuit une analyse de plus en plus fine non pas en fonction d'un plan conçu au préalable mais en fonctions des résultats obtenus à chaque étape.

2. Les calculs statistiques

Les sous-routines statistiques ne sont pas des compléments obligés des modules de comptage et d'impression de tableaux. Elles constituent, en quelque sorte, la partie facultative de la fonction calcul et dépendent avant tout de la nature même de la source ainsi que du projet de recherche et des méthodes d'interprétation. Le tableau de fréquence peut être, en effet, considéré pour son intérêt

descriptif propre ou fournir la matière première à une analyse statistique. C'est pourquoi il est légitime de ranger parmi les calculs statistiques aussi bien les calculs de pourcentage et les *ratios* simples (taux d'activité, taux de sénilité, etc. ...) que les calculs de caractéristiques plus complexes.

Cette distinction n'a rien d'artificiel car elle implique une attitude différente face aux problèmes des sous-routines statistiques. Il faut savoir, en effet, qu'il n'existe pas de fonction PL/1 qui, par elle-même, calcule automatiquement tel *ratio* ou tel paramètre statistique. Les fonctions effectuent telle ou telle opération arithmétique mais elles sont intégrées à un ensemble d'instructions qui, seules, permettent à l'ordinateur de fournir le résultat recherché. En un mot, les calculs statistiques doivent être programmés au même titre que les comptages ou l'impression des tableaux de fréquences. Dès lors plus les formules seront complexes plus la programmation le sera. Ce corollaire a pour conséquence de poser le problème de la rentabilité d'une sous-routine statistique intégrée aux modules de la fonction calcul. La programmation d'un test comme le X^2 ou le coefficient de corrélation entraîne une occupation de mémoire centrale importante, tant en place qu'en temps de travail en raison de la longueur des opérations arithmétiques et de leur nature, nous avons renoncé à introduire des sous-routines spécifiques dans les programmes de calcul . Tout d'abord en raison du coût additionnel de ces sous-routines (39); ensuite parce que les calculs statistiques ont, pour la plupart, un caractère facultatif et que l'intérêt de leur utilisation est fonction du résultat des comptages. Prévoir *a priori* plusieurs sous-routines de calcul revient à prendre le risque de faire travailler l'ordinateur en pure perte car l'importance ou la répartition des effectifs peut rendre inutile pareil traitement.

b. Les tableaux

Les sous-routines d'impression de tableau relèvent tout autant de la fonction d'édition que de la fonction calcul. Elles constituent, en effet, une forme de sortie de données au même titre que les listes dont nous avons parlé précédemment. Si nous les avons rangées dans la fonction calcul c'est qu'elles concernent, en ordre principal, des valeurs numériques et qu'elles apparaissent comme le corollaire des opérations de comptage. La notion de tableau à laquelle nous nous référons constamment doit être précisée. Nous entendons par là tant le simple relevé de fréquences d'une occurrence que le tableau statistique proprement dit ou tableau croisé qui combine deux ou plusieurs variables. Bien que l'on puisse concevoir, sur le plan théorique, des tableaux à *n* dimensions, nous nous bornerons, quant à nous, aux cas de tableaux à deux dimensions puisque ce sont les plus courants et les seuls auxquels nous ayons eu recours.

1. *Les tables*

   Les tables, ou tableaux à une dimension, sont utilisées lorsqu'il s'agit de mesurer la fréquence d'une variable discrète ou continue. Leur usage toutefois n'est pas réservé aux valeurs numériques et l'on se sert également des tables dans l'organisation de données alphabétiques pour autant qu'elles soient de même type.

2. *Les tableaux à deux dimensions*

   Nous avons vu précédemment comment se constituait un tableau de fréquences à deux dimensions

où chaque compteur était défini par des coordonnées, en l'occurrence les valeurs attribuées aux indices de lignes et de colonnes. L'impression d'un tableau semblable fait appel au même procédé de programmation que celui utilisé pour les tables.

Toutefois le fait même qu'il s'agisse d'un tableau à deux dimensions, rend la sous-routine d'impression plus complexe. D'autant qu'il est nécessaire, pour la bonne compréhension, de préciser la nature des paramètres «ligne» et des paramètres «colonne».

C. CONCLUSIONS

LEGIA constitue, à notre connaissance, le seul système complet qui effectue le traitement automatique des listes nominatives de recensement. Nous appelons complet un système qui prend en charge, à la fois et successivement, l'organisation préalable des enregistrements, le traitement de l'information selon les schéma d'analyse retenus et les sortes des résultats dans les formes souhaitées. Un tel ensemble suppose une succession de procédures, autonomes et complémentaires, qui s'intègrent les unes aux autres comme autant de modules. Elles ont chacune leur logique propre qui s'inscrit dans la logique générale dont s'inspire le système.

LEGIA se compose de trois procédures principales : la translation, la fonction d'édition et la fonction calcul. Si la première d'entre elles assume exclusivement l'organisation des *records*, les deux autres, en revanche, comprennent, chacune pour leur part, des programmes de traitement des données et des routines d'impression qui s'exécutent à partir d'un choix préalable de variables exprimé dans les critères de tri.

Ainsi conçu, le systèle *LEGIA* présente le double avantage de conserver les données originales dans leur intégrité et de permettre des exploitations «à la carte». Les seules limites apportées au choix de l'utilisateur sont celles qu'il se fixe lui-même en fonction des buts poursuivis par la recherche, de la valeur des données et de l'intelligibilité des résultats demandés. L'historien et le démographe peuvent dès lors, par l'application de ce système, obtenir tous les renseignements et les informations qu'ils désirent tant au niveau des groupes, définis selon les critères les plus divers, qu'au niveau des individus eux-mêmes. *LEGIA* leur fournit. au moindre coût, avec une égale facilité, une liste nominative des tisserands nés à Verviers et recensés dans cette ville, un tableau de répartition, de la population féminine suivant l'âge et l'état matrimonial, un relevé statistique des statuts sociaux dans le secteur textile, etc ... Les extraits de listes et les tableaux qui figurent en annexe donnent un aperçu des immenses possibilités offertes.

Soumis à ce système, les fichiers réagissent commen autant de banques de données qu'il est loisible de consulter soit par les procédures *batch* (ce sont celles décrites dans ce chapitre) soit en *time sharing* (TSO). Un profil d'informations peut être ainsi dessiné par l'utilisateur puis précisé et affiné au fil des questions posées à l'ordinateur. Si le coût de l'investigation en mode conversationnel (TSO) est trop élevé, on procède alors, comme nous l'avons fait, à un choix préalable de variables hiérarchisées et on modifie, selon les besoins, les paramètres des programmes.

Sans doute les férus des *Packages* traditionnels, S.P.S.S., OSIRIS et autres, reprocheront-ils au système *LEGIA* l'absence de procédures de calculs statistiques autres que les pourcentages et les

ratios simples. La réponse à cette critique se trouve dans la définition même des objectifs de *LEGIA*. Contrairement à S.P.S.S. ou à OSIRIS, notre système ne vise pas à analyser quantitativement des données contenues dans le fichier mais plutôt au traitement de l'information pris dans son sens le plus large, y compris les opérations statistiques élémentaires. *LEGIA* est un instrument d'investigation volontairement souple et diversifié qui s'adapte aux données alors que les *Packages*, au contraire, ont des fonctions définies une fois pour toutes qui s'appliquent à un schéma d'enregistrement déterminé. Si ces derniers sont coûteux mais hautement performants au départ de données numériques de format fixe, le système *LEGIA* présente l'avantage d'être infiniment moins coûteux et d'effectuer les traitements souhaités sans aucune contrainte sur le plan des enregistrements. Ils ne sont d'ailleurs pas exclusifs l'un de l'autre dans la mesure où l'utilisateur qui désire une analyse statistique élaborée peut soumettre à S.P.S.S. ou à OSIRIS les résultats chiffrés des traitements opérés par *LEGIA*.

Notre système veut répondre avant tout aux désidérata des historiens en leur fournissant le moyen de «manipuler», dans le bon sens du terme, la documentation enregistrée sur support informatique comme ils le feraient, *mutatis mutandis,* des fiches traditionnelles. Avec l'apport décisif de la vitesse d'exécution et de l'inépuisable capacité de travail qui sont l'apanage incontesté de l'ordinateur, *LEGIA* apporte, au problème posé par l'exploitation de sources massives, une solution décisive.

****