Les procédures et les moyens informatiques pour l'enquête sur les 3000 familles "TRA"

Jean-Claude Poupa

Institut National de la Recherche Agronomique

Station d'Economie et Sociologie Rurales
65, rue de St-Brieuc - 35042 Rennes cedex

LES PROCÉDURES ET LES MOYENS INFORMATIQUES POUR L'ENQUÊTE SUR LES 3 000 FAMILLES "TRA"

Octobre 1993

Jean-Claude Poupa

L'INRA, avec l'accord du CNRS, a confié à Jean-Pierre Pélissier une mission de proposition [1] concernant l'organisation scientifique et matérielle nécessaire à la poursuite de l'enquête sur la mobilité géographique et sociale en France au XIX^{ème} et XX^{ème} siècle, lancée en 1980 par Jacques Dupâquier. L'objectif principal de cette opération est "la réalisation d'une base de données informatique qui devra pouvoir être mise à disposition de la communauté scientifique". Les propositions devront porter sur "la conception scientifique et technique de la base de données, les modalités d'organisation de l'équipe de projet".

Ce rapport définit un cadre théorique pour la recherche et la production de solutions informatiques adaptées à la taille et à la complexité de l'enquête. Il propose un scénario d'exécution basé sur une première modélisation logique élaborée à l'issue d'une reconnaissance du contenu des fichiers informatiques disponibles. Cette modélisation est décrite de façon relativement détaillée en s'appuyant sur des exemples simples. L'objectif n'est pas de soumettre une solution "clés en main", mais de montrer l'existence d'instruments mathématiques qui pourraient être efficaces pour résoudre un certain nombre de problèmes.

Un premier instrument est le calcul relationnel, tel qu'il a été défini par CODD en 1970. Son efficacité théorique et pratique a pu être montrée à l'INRA pour le traitement des données de panels nationaux. Un second instrument, qui reste à expérimenter dans ce type d'enquête, est issu de la théorie des langages de CHOMSKY[2]. Il utilise les notions de grammaire formelles et d'automates finis pour traiter les variations syntaxiques et les transcriptions phonétiques qui interviennent dans le procéssus d'établissement des documents d'état civil. La construction de cette grammaire est une opération complexe qui passe par la formalisation et l'intégration d'un ensemble de règles qui reproduisent le savoir-faire du généalogiste.

---

[1] Lettre de mission du 21/04/93.

[2] Une présentation synthétique de ce concept est faite dans "Le courrier du CNRS" consacré à la recherche en informatique (n°80) Février 1993.

1. LE CONTENU DE LA BASE

L'objectif est de rassembler dans une base de données des informations relatives aux individus qui descendent d'un échantillon initial de 3 000 couples mariés entre 1803 et 1832, représentatif de la population française de l'époque (recensement de 1806). L'échantillon a été choisi parmi les familles dont les patronymes commencent par les trois lettres "TRA". L'enquête ne retient que les individus qui portent ce patronyme. Les informations démographiques sont extraites des actes de l'état-civil (mariage, naissance, décès) relatifs aux individus de l'échantillon final (3 000 TRA et tous leurs descendants porteurs de ce patronyme initial avec prise en compte des variantes de transcriptions). Elles doivent être complétées par des informations relatives aux patrimoines déclarés lors des successions (tables de successions et absences des centres des impôts).

L'expression "base de données" désigne ici un *ensemble* structuré d'informations accessibles au moyen de méthodes de calcul logique pour l'équipe de recherche responsable de l'enquête. Ces bases sont privées et des mesures de sécurité doivent être mises en place pour interdire les accès externes, conformément aux dispositions légales relatives aux données individuelles.

1.1. La genèse de l'échantillon

Un point fondamental est la constitution de l'échantillon de tous les individus issus des 3 000 familles initiales, défini sur une période de près de deux siècles, et porteurs d'un patronyme préfixé par "TRA". Un individu *i* appartient à l'échantillon E si son père légal fait partie de l'échantillon. La même règle s'applique pour les parents, jusqu'à remonter à la famille souche de la période initiale. Pratiquement, cela signifie qu'il faut rechercher pour tout individu TRA des informations pour identifier ses parents. La recherche s'effectue de génération en génération jusqu'à la période initiale, le couple parent appartenant, ou n'appartenant pas, à l'ensemble des 3 000 familles retenues. Pour conclure, il faut donc connaître l'identité de tous les couples TRA de la période de départ (1803-1832), soit un ensemble d'au moins 11 000 couples dans l'état actuel des fichiers disponibles.

Ce travail préliminaire doit aboutir à la production d'une première base contenant pour chaque individu TRA, quelle que soit la période, une information minimale pour une identification non ambiguë (ce point sera abordé plus loin) et une marque précisant s'il appartient ou non à l'échantillon. La mise à jour se fera dynamiquement, un individu étant "marqué" en recherchant la personne qui a transmis le patronyme.

Au terme d'une approximation sommaire, cette base devrait regrouper plus de 100 000 individus, dont la moitié environ appartiendrait à l'échantillon. Un moyen de réduire la taille de cette base serait de se restreindre à un sous-ensemble de patronymes sur la période initiale.

1.2. Les données démographiques et patrimoniales

Les informations détaillées contenues dans un document sont à extraire seulement si elles concernent un individu de l'échantillon. En l'absence de la base précédente, il faut saisir et valider tous les documents dans lesquels émargent des individus TRA, et vérifier ultérieurement l'appartenance à l'échantillon. Cette seconde solution aboutit à constituer de fait un échantillon sur la totalité des TRA : elle a été mise en oeuvre pour les actes de mariage du XIX^{ème} siècle et les actes de naissance du XX^{ème} siècle disponibles.

2. GESTION DE L'ECHANTILLON DES INDIVIDUS

La génération de l'échantillon est une étape préliminaire fondamentale. Il est indispensable d'identifier de façon unique tous les individus susceptibles d'appartenir à l'échantillon et de représenter les liens de filiation entre ces personnes physiques. Ces filiations peuvent être vues comme des relations binaires du type *"x est enfant de y"* sur l'ensemble des personnes dont le nom est préfixé par "TRA" et ayant vécu entre 1803 et maintenant. La structure de données associée est un ensemble d'éléments sur lesquels sont définies des relations binaires, lesquelles traduisent formellement les liens usuels de parenté.

2.1. L'identification des personnes

Une identification non ambiguë des personnes nécessite la prise en compte d'attributs propres évidemment invariants dans le temps : les lieux de résidence sont ainsi à éliminer. Les noms et prénoms propres subissent des variations au niveau de la transcription, qui se traduisent par la définition d'équivalences (consonnes doubles, phonèmes proches...). Enfin, il y a lieu de choisir des informations qui sont présentes dans la plupart des documents.

Nous avons examiné, pour les 45 000 actes de mariage du XIX^{ème} siècle, l'efficacité d'un identifiant formé des noms et prénoms usuels des époux (avec prise en considération de deux prénoms si le premier est "Jean" ou "Marie"). Après élimination d'actes probablement saisis en double (même date et même lieu), il reste moins de 100 identifiants qui désignent deux actes. L'examen de ces actes est à faire pour déterminer le nombre exact d'ambiguïtés vraies, pour lesquelles l'identifiant désigne effectivement des personnes physiques différentes. Une telle ambiguïté est illustrée sur l'exemple de deux couples logiquement indiscernables (Joseph Trabichet, Françoise Favre), mariés respectivement en 1807 et 1816 à Vailly en Haute-Savoie : on dénombre 7 mariages d'enfants de prénoms différents issus de ces deux couples, sans qu'il soit immédiatement possible, au vu des actes, d'affecter les enfants à une famille.

La prise en compte des variations dans l'écriture des noms propres et prénoms risque d'introduire de nouvelles ambiguïtés. Cette question est complexe et nécessite la définition d'équivalences, respectivement sur les ensembles de noms propres et de prénoms (on dira par exemple que "Michèle" appartient à la même classe que "Michelle"). Ce volet est traité au chapitre 6.

Nous retiendrons cependant de ce premier examen que les noms et prénoms usuels des époux, en tenant compte d'une variabilité des transcriptions, identifient la plupart du temps (vraisemblablement dans plus de 99 % des cas) une famille unique. Quand il y a ambiguïté, elle porte sur l'ensemble des mariages du XIX^{ème} siècle sur deux couples seulement (nous avons néanmoins repéré quatre triplets qui correspondent à des actes saisis trois fois). On notera qu'une ambiguïté portant sur trois couples ou plus est peu probable mais possible.

La famille étant identifiée, tout individu né dans cette famille est alors repéré par son prénom, sous réserve que deux enfants issus d'un même couple n'aient pas un prénom identique (ce qui reste possible).

2.2. Les relations de filiation entre personnes

L'échantillon initial de "TRA" de la première génération étant choisi, il s'agit de suivre les générations successives jusqu'à la génération actuelle. La difficulté pratique est de retrouver les conjoints des individus mariés pour suivre la descendance, un même individu pouvant se marier plusieurs fois.

2.2.1. A partir des actes de mariage

Les actes de mariage contiennent l'information nécessaire pour construire dynamiquement l'échantillon final, les couples parents étant identifiés. Cette démarche suppose que l'on progresse par générations successives, un couple de la génération *h* appartenant à l'échantillon si les parents de la génération (*h-1*) appartiennent à l'échantillon.

La méthode permet de recueillir l'information sur les filles "TRA" mariées (sans suivre leur descendance sauf si l'époux est lui aussi un "TRA" de l'échantillon), mais ignore les célibataires et la descendance des mères célibataires.

2.2.2. A partir des actes de naissance

Pour le XX^{ème} siècle, cet acte permet de construire des filiations partielles dès lors que l'identité du conjoint figure normalement en mention marginale ajoutée à la date du mariage. Les branches devront être raccordées à une génération du XIX^{ème} siècle au vu d'un acte de mariage pour déterminer l'appartenance à l'échantillon.

2.2.3. A partir des actes de décès

L'acte de décès, s'il est complet, est un document suffisant pour construire les filiations. A notre connaissance, il n'a pas été exploité.

2.3. L'extraction de l'échantillon

Les actes sont recherchés après consultation des tables décennales, insuffisantes pour identifier les personnes. Le relevé peut se faire de façon quasi-exhaustive pour les actes de naissance et de décès, classés dans l'ordre alphabétique. En revanche, l'inventaire des mariages des filles TRA mariées nécessite de parcourir les tables décennales du début à la fin, le classement étant fait en fonction du nom de l'époux : le risque d'oubli est élevé.

2.4. Les liens avec les sources d'informations

Tout acte d'état civil peut être repéré par un numéro unique, fixé arbitrairement. Pour des actes transcrits dans des fichiers séquentiels sous forme d'enregistrements de format fixe, il suffit de prendre le numéro d'ordre *(na)* dans ce fichier. Au sein d'un enregistrement, un individu peut être repéré par son numéro en séquence *(ni)*. Cet individu est identifié par son prénom *(iprenom)*, les nom *(pnom)* et prénom *(pprenom)* de son père, les nom *(mnom)* et prénom *(mprenom)* de sa mère. Sur l'ensemble des actes de mariage, un individu serait désigné de façon unique par le 7-uplet *(na, ni, iprenom, pnom, pprenom, mnom, mprenom)*, le problème des transcriptions phonétiques étant traité plus loin.

Avec l'exemple du couple Trabichet-Favre cité plus haut, la fille Julie, mariée à Vailly en 1843, est représentée par la valeur (31 364, 1, Julie, Trabichet, Joseph, Favre, Françoise). Cette notation signifie que Julie Trabichet est citée comme individu numéro 1 dans l'acte de mariage numéro 31 364. On notera que le quintuplet (Julie, Trabichet, Joseph, Favre, Françoise) peut apparaître dans plusieurs actes, et désigner des personnes physiques différentes. Cette éventualité, possible mais peu probable, doit néanmoins être prévue. Le couple (31 364, 1) désigne quant à lui une personne et une seule.

Pour poursuivre l'exploitation de cet exemple, on peut souligner que si l'identifiant (Favre, Françoise) est ambigu, les identifiants (Françoise, Favre, Aimé, Grillard, Marguerite) et (Françoise, Favre, Claude, Verney, Marie) paraissent désigner exclusivement les deux personnes ayant effectivement vécu à Vailly en Haute-Savoie au début du XIX^{ème} siècle.

3. LE CONTENU DES FICHIERS INFORMATIQUES

L'observation élémentaire dans cette enquête est l'individu qui a vécu sur une période donnée. Un même individu est cité dans plusieurs actes, et possède un rôle dans chaque acte (nouveau né, marié, défunt, parent, conjoint...). Les valeurs observées concernent des variables ou attributs propres aux personnes citées, à définir. L'objectif de cette enquête est de construire ces ensembles d'individus, d'établir des relations sur ces ensembles (l'élément x représente la même personne que l'élément y, x est le père de y...), puis de repérer des valeurs pour une liste de variables préalablement définies.

Nous disposons pour cela d'une collection de fichiers non documentés, pour lesquels il faut réaliser une reconnaissance syntaxique et sémantique avec des outils d'analyse à construire au vu des données. A l'issue d'une première exploration, il est possible de décrire plus précisément la nature des informations disponibles sur support informatique.

3.1. Les tables décennales

Les tables décennales contiennent les informations utiles pour localiser les actes de naissance, mariage et décès : lieu (département et commune), date (jour, mois, année), nom et prénoms de l'individu sujet de l'acte, nom et prénoms du conjoint pour les mariages.

Ces informations sont codées dans des fichiers "texte", avec des formats d'enregistrements différents selon les fichiers. Certains champs occupent une position et une longueur constante : lieu, date, type d'acte. La valeur peut alors être extraite en isolant une sous-chaîne dans l'enregistrement (année en position 38-41 par exemple). Le champ d'identification de l'individu, bien que codé dans une sous-chaîne de longueur fixe (position 44-100), doit être interprété pour extraire le nom propre et une liste de prénoms éventuellement tronquée. Ce champ peut être décrit sous la forme de Backus et Naur au moyen de la règle syntaxique suivante :

<individu> :: = <nom>, [<prenom>] {<prenom>}.

Cette notation, couramment utilisée pour la description syntaxique des langages, signifie ici que le composant <individu> est formé de plusieurs composants, le premier suivi nécessairement du caractère ", ", les autres optionnels et séparés par un espace. Si l'on souhaite introduire un contrôle plus strict, on ajoutera une règle précisant que les composants <nom> et <prenom> sont formés par concaténation des lettres de l'alphabet. Il faudra également prévoir des cas particuliers liés à des conventions (présence d'un trait d'union, d'une abréviation...). Enfin, pour les mariages, il est nécessaire de décoder deux identifiants de personnes, situés de part et d'autre du caractère " = ". Cet exemple est cité pour introduire la nécessité de recourir

à des méthodes et techniques de compilation [3].

Des informations complémentaires, en particulier le nom propre TRA "normalisé", sont ajoutées dans certains fichiers. La fonction principale de ces documents est cependant de dénombrer et localiser les autres. Ils fournissent des "pointeurs" sur des actes qui contiennent des informations relatives aux individus susceptibles d'appartenir à l'échantillon, et constituent un instrument pour extraire cet échantillon.

Une première tâche informatique est l'analyse des différents fichiers relatifs aux tables décennales, avec apuration et élimination des doublons pour calculer trois ensembles disjoints : naissance, mariage, décès. La cardinalité de ces ensembles se mesure en centaines de milliers. Le "pointage" sur les actes à dépouiller se fait au moyen du triplet *(lieu, date, nom)*. Pour les mariages, le nom choisi est celui du mari s'il faut extraire les actes relatifs aux mariages des filles "TRA".

3.2. Les actes de mariage

Le fichier examiné regroupe un ensemble de plus de 45 000 enregistrements relatifs pour l'essentiel à des actes du XIX^{ème} siècle (96 %). Le début de l'enregistrement est une entête précisant les lieux, dates et numéros d'actes. Le premier caractère de l'entête est un chiffre qui indique si l'acte de mariage concerne un fils TRA (1) ou une fille TRA (2).

L'entête est suivie systématiquement d'un ensemble de six segments contenant des informations sur des personnes citées dans l'acte. Les trois premiers segments concernent successivement l'époux porteur du patronyme TRA (homme ou femme) et ses parents. Les trois segments suivants sont relatifs au conjoint et à ses parents. Cette convention fait que l'ordre des personnes dans l'acte initial n'est plus respecté pour le mariage d'une fille TRA, sauf si son époux porte le patronyme TRA. Ce point est source ambiguïté et il paraît prudent de rétablir l'ordre de l'état civil pour coder le rôle de la personne dans l'acte. On remarquera que parmi les 491 mariages entre TRA, 442 sont rattachés à un fils (1) et 40 à une fille (2).

Si l'on appelle segment non vide une zone dans un enregistrement de longueur et position fixes et dont le premier caractère est une lettre de l'alphabet (pratiquement la première lettre d'un nom propre), l'exploration du fichier des mariages fournit des résultats résumés et interprétés dans le tableau 1. Les segments 7 et 8 n'apparaissent qu'épisodiquement [4].

---

[3] La traduction est faite en langage C.
[4] Il pourrait s'agir des anciens conjoints.

Tableau 1. Les segments "individus" dans l'enregistrement "acte de mariage"

Numéro ordre	Rôle de l'individu	Position segment	Longueur segment	Nombre de segments
1	mari ou femme TRA	20	144	45 071
2	père du TRA	164	118	44 006
3	mère du TRA	282	118	44 105
4	conjoint du TRA	400	145	45 034
5	beau-père du TRA	545	118	43 881
6	belle-mère du TRA	663	117	44 214
7	indéterminé	780	134	4 409
8	indéterminé	914	94	2 686

Les six premiers segments contiennent des dates, lieux (codés et en clair) ainsi que le métier exercé. Ils se terminent généralement par les lettres "S" ou "N" (code signature). Les dénombrements traduisent l'existence d'erreurs syntaxiques, marginalement de patronymes non transcrits. Une erreur a réduit de 1 caractère le segment numéro 6 : le code numérique du lieu est tronqué et se trouve réduit à 4 chiffres (au lieu de 5).

L'édition de ces segments avec une police à espacement constant et une taille de caractères adaptée permet de repérer les champs et d'avancer des hypothèses quant à la structuration syntaxique [5]. Les dates et lieux occupent en général une position constante, à des variations marginales près qui peuvent traduire des erreurs de saisie. L'identification de la personne, sous la forme d'un nom suivi d'une liste de prénoms, est codée dans un composant situé au début du segment et dont la largeur est restreinte à 40 caractères. Ce composant syntaxique est présent au début des segments de rang 7 et 8, lorsqu'ils ne sont pas vides.

La fin d'un enregistrement paraît contenir le nom TRA dit "normalisé" en position 1008-1022. La plupart des enregistrements ont une longueur fixe de 1 040 caractères : les informations figurant au delà de la position 1022 n'ont pas été interprétées.

Ces hypothèses relatives à la structure syntaxique de l'enregistrement restent à valider au cours d'une phase d'analyse qui devrait déboucher sur la formalisation et la traduction des algorithmes de reconnaissance et de contrôle adaptés.

Les individus cités dans les actes peuvent apparaître plusieurs fois, comme époux ou parents lors des mariages des enfants. Les fichiers fournissent également des renseignements sur des individus explicitement hors descendance, les beaux-parents des TRA.

---

[5] Ces éditions utilisent le langage POSTSCRIPT.

3.3. Les actes de naissance

Le fichier examiné regroupe un ensemble de 42 000 enregistrements relatifs à des actes de naissance du XX^{ème} siècle. Un enregistrement peut être décomposé en segments, repérables par leur position et leur longueur. L'entête (34 caractères) précise la date et le lieu. Le segment suivant (52 caractères) regroupe les informations relatives au nouveau-né. Les informations sur les parents sont codées dans deux segments de structures apparemment identiques (148 caractères par segment) : identité, date, âge, lieu, métier.

Pour 25 000 actes figure un segment relatif au mariage de l'enfant (99 caractères) : identité de l'époux, lieu et date de l'acte. Pour un peu moins de 7 000 actes existe une référence au décès : lieu et date. Les informations codées entre les positions 517 et 778 n'ont pas été interprétées. Le nom normalisé du TRA est reproduit en fin d'enregistrement (779-794).

Les individus sont repérés par leurs noms et prénoms, séparés par des espaces (il n'y a pas de délimiteur spécifique pour séparer le nom propre du prénom).Ils peuvent apparaître plusieurs fois, comme nouveau-né puis parents.

La référence au mariage permet de "pointer" sur des actes de mariage dans les tables décennales au moyen du triplet *(date, lieu, nom)*. Le nom choisi sera celui du mari, pratiquement celui du nouveau-né pour les hommes ou celui figurant dans la référence au mariage pour les femmes.

3.4. Les autres fichiers

Le fichier des personnes physiques de l'INSEE fournit des renseignements sur l'identité des personnes, les dates et lieux de naissances et décès. L'utilisation de ce fichier dépend de la stratégie qui sera adoptée pour générer l'échantillon final : c'est un moyen de vérifier l'exhaustivité des relevés des tables décennales pour les naissances et les décès.

Les tables de successions et absences ne sont utilisées dans un premier temps que pour rechercher des individus TRA qui auraient été oubliés dans les autres relevés (tables décennales non dépouillées).

4. LES METHODES INFORMATIQUES

Cette enquête a démarré au début des années 80 : son exploitation informatique n'était à l'époque envisageable que sur les centres de calcul des institutions de recherche. L'information, saisie "au kilomètre" par des opérateurs spécialisés, est restituée dans des fichiers séquentiels regroupant des enregistrements logiques décrits par un format. La phase d'apuration consiste à traduire dans un langage de programmation (fortran, cobol, pl1) des tests logiques pour détecter les erreurs et incohérences, puis réaliser les corrections. Les grandes institutions statistiques ont mis en place des systèmes adaptés au dépouillement d'enquêtes, depuis la saisie jusqu'à la restitution des résultats apurés, sous la forme de tableaux éventuellement repris par des logiciels statistiques. Un exemple classique est le logiciel LEDA de l'INSEE. Ces méthodes centralisées mobilisent en fait beaucoup de ressources informatiques, logicielles et humaines.

Pour l'enquête TRA, il semblerait que le processus se soit arrêté à la saisie, le rapport Villac évoquant l'existence "de programmes maison, écrits en pl1 ou en SAS, sur la qualité desquels existent les plus grands doutes". La phase d'apuration est donc à réaliser, en tenant compte du fait que les données ont par ailleurs pu être altérées.

Le point de départ de la chaîne informatique de traitement est finalement un ensemble complexe d'enregistrements logiques, de sources multiples, et dont les plus récents peuvent être issus de logiciels de saisie spécialisés fonctionnant sur microordinateurs. La structure syntaxique de ces enregistrements est à reconnaître, en l'absence de toute documentation.

4.1. La reconnaissance syntaxique des enregistrements

L'objectif de cette étape est d'aboutir à un système dans lequel l'équipe de recherche est en mesure de rechercher et manipuler l'information globalement sans devoir se soucier de la façon dont le système informatique code cette information et la gère dans les fichiers.

Le principe est de répartir dans des **classes** les **segments** qui composent les enregistrements des différents fichiers. Ces segments correspondent à des zones bien localisées de l'enregistrement, qui regroupent un ensemble d'informations relatives à une entité, la plupart du temps un individu ayant un rôle dans un acte. Ils doivent être identifiés au moyen de règles et codes instrumentaux (numéro du segment, numéro d'enregistrement, fichier source) pour pouvoir recomposer intégralement l'information initiale. Pratiquement, les segments sont d'un point de vue informatique des tableaux de caractères dont la taille varie de 10 à 200 éléments. Une classe regroupe des entités de même nature qui peuvent provenir de fichiers différents pourvu que la source soit identifiée.

Ces segments sont gérés dans un système de gestion de **bases de données relationnelles** (SGBDR). Un segment constitue dans ce cadre formel un élément indivisible, ou **donnée atomique**. Il est toutefois interprétable par un expert qui lit la chaîne de caractères associée, reconnaît des variables et affecte une sémantique à ces variables. Cette démarche permet de construire progressivement des algorithmes de reconnaissance syntaxique, traduits et expérimentés sur des échantillons. Un exemple d'algorithme simple est celui de l'extraction du nom propre et du prénom usuel.

Le chargement de la base nécessite la définition préalable d'algorithmes de décomposition des enregistrements en segments, avec un mécanisme de désignation non ambiguë des éléments. Les contrôles sont réduits au minimum dans cette étape pour restreindre le nombre de rejets. Cette base, dite **archive**, est construite pour apurer l'information au moyen des outils disponibles autour du SGBDR, dont le langage SQL.

4.2. Les types de données abstraits

Le besoin de conserver et visualiser l'information initiale, la structure syntaxique étant pour l'instant inconnue, conduit à stocker les segments sous une forme textuelle. Au sein d'un système relationnel, un texte est un élément atomique, qui appartient à un ensemble de définition appelé **domaine**. Le langage SQL intègre les opérateurs arithmétiques sur des ensembles de nombres, les opérateurs et prédicats ensemblistes ainsi que l'opérateur de concaténation sur l'ensemble des chaînes de caractères. Les SGBDR fournissent généralement des fonctions de traitement de chaînes de caractères qui permettent d'extraire des sous-chaînes : ces extensions au langage SQL ne sont pas normalisées et la syntaxe varie d'un système à l'autre (substr, right, left...). L'utilisation de ces fonctions "propriétaires" lie en fait l'application SQL à un logiciel spécifique. Les fonctions de traitement des chaînes de caractères sont cependant utiles pour extraire des champs de position et longueur constantes, ou trouver une occurrence d'un nom commençant par le préfixe "TRA". Elles sont inadaptées dès qu'il faut analyser un composant syntaxique plus complexe, voire dangereuses lorsqu'au terme d'une accumulation "d'astuces" il n'est plus possible de démontrer l'exactitude du résultat.

En l'absence de ces fonctions, nous proposons de les construire pour les traduire dans un langage dûment normalisé. En reprenant le vocabulaire de la programmation orientée objet, ces fonctions sont appelées **services** (niveau logique) ou **primitives** (niveau exécution). La description sera faite au moyen de **grammaires** pour les composants syntaxiques complexes. La structure de données et les fonctions définies sur cette structure constituent un **type abstrait de données.**

La fonction qui reconnaît le nom propre et régénère la liste des prénoms (avec reconnaissance et extension des abréviations) est un exemple type de service qui devra être validé, et restera associé à la base archive pendant la durée de vie du projet. Une fonction d'édition d'une chaîne de caractères qui admet comme paramètres un nom de police et une taille constitue un autre service associé à la structure de données.

Dès lors que les services sont définis sur une structure de données, il est intéressant de pouvoir stocker les primitives dans la base, et de demander l'exécution pour des classes d'éléments qui utilisent ces structures [6]. A défaut, les opérations sont exécutées à l'extérieur de la base sur des copies fichiers des éléments à traiter.

4.3. L'identification des entités ou segments

Une identification non ambiguë est de désigner le $j^{ième}$ individu du $i^{ième}$ acte d'une source *s* par le triplet *(s, i, j)*, dit **clé instrumentale**. Cette convention, qui exprime la notation indiciaire usuelle, est efficace pour dialoguer avec le système mais ne fournit pas aux utilisateurs les fonctionnalités attendues. Les identifiants utilisés par les experts sont des patronymes et des prénoms pour des actes établis en un lieu et à une date donnés. Nous appellerons **clé sémantique** le n-uplet utilisé par l'expert pour rechercher les informations relatives aux individus. Une famille est identifiée par le nom du mari *(mnom)*, le prénom du mari *(mprenom)*, le nom de la femme *(fnom)*, le prénom de la femme *(fprenom)*. Un enfant est désigné par son prénom *(eprenom)* et l'identité de sa famille. Un acte est repéré par l'année *(date)* et le lieu *(lieu)*. La clé sémantique associée à un individu est alors le 7-uplet :

(eprenom, mnom, mprenom, fnom, fprenom, date, lieu).

Cette clé d'expert peut en théorie désigner plusieurs individus. Sur la base de l'exemple Trabichet-Favre cité plus haut, il suffit que naissent dans chacune des familles pour une année donnée des enfants qui porteraient le même prénom. Une telle clé n'est donc pas d'un point de vue théorique utilisable (elle ne désigne pas nécessairement un élément et un seul), mais s'avère être indispensable sur le plan pratique.

La construction de cette clé sémantique nécessite d'extraire des enregistrements les informations d'identification, ce qui revient à construire et valider à l'extérieur du SGBDR les premières primitives associées aux segments. L'exécution de ces fonctions va conduire au rejet de certains actes suite à la violation d'une propriété : nom propre ne commençant pas par une

---

[6] Les types abstraits de données sont gérés dans certains SGBDR pour lesquels existent des extensions du modèle relationnel dites "objets".

lettre, code lieu non numérique, date erronée... . Les actes rejetés sont à apurer en partie dans le système de gestion de fichiers pour être ensuite chargés dans la base. A ce stade les contrôles doivent être aussi restreints que possibles, l'apuration proprement dite se poursuivant dans la base : appartenance aux dictionnaires des patronymes, prénoms, lieux, etc... . L'objectif recherché est en fait de passer le plus rapidement possible des fichiers au modèle relationnel.

4.4. Gestion des règles et cheminement dans un graphe

Un élément de l'ensemble des TRA représente un individu, ayant vécu sur la période d'observation et porteur du patronyme TRA. L'enquête s'intéresse aux lignées qui ont transmis ces patronymes. Soit deux éléments x et y. Le problème se formalise simplement en théorie des graphes en définissant un **arc** du sommet x vers le sommet y, lequel exprime la relation "x descend de y". Cet arc traduit le lien "enfant —> père". La méthode de génération de l'échantillon fait que d'un sommet part au plus un arc vers l'individu ayant transmis le patronyme TRA : la relation "enfant —> mère" ne doit pas être représentée pour une épouse portant le patronyme TRA et mariée avec un TRA.

Ce graphe est sans circuit.

Une lignée en descendance paternelle est, au sens généalogique, une **arborescence**. La longueur d'un chemin représente un nombre de générations. Les individus de la première génération sont des racines d'arborescences.

Soit *père (x)*, la fonction qui désigne l'individu qui a transmis le patronyme, ou une valeur négative si cet individu est inconnu. L'énumération des lignées s'effectue alors en appliquant, sur tous les sommets sur lesquels n'arrivent aucun arc, la règle récursive suivante :

si *père (x)* < 0 alors arrêt ; sinon père (père (x)) ;

L'application de cette règle ne serait plus possible si les ascendances paternelles et maternelles étaient représentées simultanément : il faudrait recourir aux algorithmes classiques de la théorie des graphes.

4.5. Profil d'un système

La solution proposée dans ce rapport s'appuie sur la théorie du modèle relationnel. Certaines définitions fondamentales sont rappelées dans le chapitre suivant.

En plus du calcul relationnel apparaît la nécessité de manipuler des types de données abstraits construits sur des tableaux de caractères. Nous n'utilisons pas la notion d'objet qui introduit d'autres concepts, inutiles dans ce contexte. L'écriture des primitives relève de l'analyse-programmation classique, une maîtrise des techniques de compilation usuelles paraissant indispensable (grammaires formelles, automates d'états finis...). Ces primitives peuvent éventuellement remplacer des opérateurs : l'exemple type est la comparaison des patronymes, l'équivalence étant testée avec l'opérateur "=".

Les règles sont formalisées à l'aide du calcul relationnel, en distinguant éventuellement les générations d'individus pour réduire la complexité.

5. ARCHITECTURE DE LA BASE D'ARCHIVAGE DES DOCUMENTS

La fonction de cette base est de regrouper l'intégralité des informations dans un modèle mathématique de données, géré au moyen de technologies adaptées. L'approche théorique est présentée dans un article annexé à ce rapport [7].

On utilisera trois **domaines** pour la définition des **attributs** :

1) l'ensemble N des entiers naturels ;
2) l'ensemble des mots, noté NOM, dont les éléments sont construits par concaténation des lettres de l'alphabet latin ;
3) l'ensemble des chaînes de caractères, noté TEXTE, dont les éléments sont construits par concaténation de caractères imprimables.

Une **relation** p-aire désigne un sous-ensemble du produit cartésien de p domaines. La notation R (i : N, x : MOT) signifie que la relation binaire R a un attribut i défini sur l'ensemble N des entiers naturels et un attribut x défini sur l'ensemble MOT. Les attributs formant la **clé primaire** sont soulignés.

Un premier groupe de relations contient les informations brutes relatives aux documents : tables décennales, mariages, naissances, etc... . Ces relations sont chargées à partir des fichiers. Un second groupe gère les liens d'identité ou de parenté entre les individus jouant un rôle dans un acte, représentés par un **tuple** ou élément de relation. Ces liens sont de fait des relations binaires : *x* est le père de *y*, *x* et *y* représentent une seule et même personne, etc... . Les relations de ce second groupe sont calculées au moyen d'opérations sur les relations du premier groupe.

5.1. L'archivage des documents

Comme indiqué plus haut, cette étape segmente les enregistrements logiques des fichiers et ajoute des attributs instrumentaux et sémantiques. Un attribut *s* (source) peut être ajouté pour distinguer des sources d'informations distinctes. Les schémas de relations proposés introduisent des redondances volontaires, dans un souci d'efficacité.

5.1.1. Les tables décennales

Les données des tables décennales sont réparties dans trois relations selon la nature de l'acte :

---

[7] J. C. Poupa. La représentation relationnelle des données statistiques : application au traitement des données de panel. Cahiers d'Economie et Sociologie Rurales, n° 26, 1993.

NAISSANCE (n : N, date : N, lieu : N, nom : MOT, prenom : MOT, x : TEXTE),
DECES (n : N, date : N, lieu : N, nom : MOT, prenom : MOT, x : TEXTE),
MARIAGE (n : N, date : N, lieu : N, Mnom : MOT, Mprenom : MOT,
Fnom : MOT, Fprenom : MOT, x : TEXTE).

L'attribut *lieu*, numérique, est clé étrangère dans le dictionnaire de définition des lieux. L'attribut *x* contient la totalité de l'enregistrement logique lu dans le fichier. La sémantique des attributs des clés est celle décrite dans le chapitre précédent. L'attribut *prénom* contient le premier prénom, ou un prénom composé si le premier est "Jean" ou "Marie".

5.1.2. Les actes de mariage

Le principe retenu est de distinguer les individus cités dans l'acte (jusqu'à huit segments). L'information est affectée dans les relations décrites par les schémas suivants :

MROLE (n : N, role : N, date : N, lieu : N, nom : MOT, prenom : MOT, x : TEXTE),
MGLOBAL (n : N, date : N, lieu : N, entête : TEXTE, reste : TEXTE).

Un acte standard contient au moins six segments, une entête et une conclusion (reste) : il génère au moins six tuples dans la relation MROLE et un tuple dans la relation MGLOBAL. L'attribut *role* est une codification du rôle de l'individu dans l'acte. L'attribut *entête* reproduit la zone d'identification de l'acte en début d'enregistrement. L'attribut *reste* reçoit les champs codés après la fin du dernier segment individu. Il peut éventuellement être éclaté en plusieurs attributs. A défaut de connaître la sémantique de l'information codée dans cette zone, une solution immédiate est de tout prendre pour différer l'analyse dans le temps.

5.1.3. Les actes de naissance

Les schémas proposés isolent les individus et prennent en compte les évolutions de l'acte avec l'ajout de la référence au mariage de l'enfant (1897) puis de la référence au décès (1945).

NENFANT (n : N, date : N, lieu : N, nom : MOT, prenom : MOT, x : TEXTE)
NPARENT (n : N, role : N, date : N, lieu : N, nom : MOT, prenom : MOT, x : TEXTE)
NMARIAGE (n : N, date : N, lieu : N, Mnom : MOT, Mprenom : MOT, Fnom : MOT,
Fprenom : MOT, x : TEXTE)
NDECES (n : N, date : N, lieu : N, nom : MOT, prenom : MOT, x : TEXTE).

L'attribut *role* de la relation NPARENT prend deux valeurs, par exemple 1 pour "père" et 2 pour "mère". Les attributs *date* et *lieu* des relations NMARIAGE et NDECES contiennent les valeurs des dates et lieux de mariage ou décès.

5.2. La gestion des liens entre individus

Les relations décrites ci-dessous sont calculées au moyen d'opérations relationnelles exprimées dans le langage SQL.

5.2.1. Le contrôle d'exhaustivité des relevés des actes de mariage

La relation d'identification des époux sujets de l'acte est construite par une opération de jointure de MROLE sur MROLE (autojointure) qui rend la relation

COUPLE (Mnom : MOT, Mprenom : MOT, Fnom : MOT, Fprenom : MOT,
lieu : N, date : N, n : N)

La syntaxe de la requête SQL qui calcule cette relation est la suivante :

```
SELECT    a.nom as Mnom, a.prenom as Mprenom,
          b.nom as Fnom, b.prenom as Fprenom,
          a.lieu, a.date, a.n
FROM      MROLE a, MROLE b
WHERE     (a.n = b.n) AND (a.role = 1) AND (b.role = 2) ;
```

Nous supposons évidemment que les valeurs respectives des codes des rôles "époux" et "épouse" sont 1 et 2. On remarquera que la clé sémantique *(Mnom, Mprenom, Fnom, Fprenom)* ne peut pas constituer une clé primaire.

La différence ensembliste exprimée sur les attributs communs des relations MARIAGE et COUPLE fournit l'ensemble des mariages cités dans les tables décennales et pour lesquels l'acte n'est pas trouvé. Inversement, la différence (COUPLE - MARIAGE) [8] fournit l'ensemble des mariages effectivement saisis mais non répertoriés dans les tables décennales. On notera qu'il ne suffit pas que ces deux différences soient vides pour prouver que tous les mariages ont été relevés, du fait de l'ambiguïté des clés sémantiques signalée plus haut.

5.2.2. La recherche des actes de mariage des parents

De la même façon que pour les époux, une autojointure sur MROLE, avec prise en compte dans le prédicat de sélection des rôles des parents rend l'identification des couples qui marient leurs enfants.

On peut ainsi construire deux relations de même schéma respectivement pour les parents du mari et de la femme, soit :

---

[8] Par souci de simplification, nous ne détaillons pas les opérations relationnelles de restriction et projection qui précèdent le calcul de la différence.

PMARI (Mnom : MOT, Mprenom : MOT, Fnom : MOT, Fprenom : MOT, n : N),
PFEMME (Mnom : MOT, Mprenom : MOT, Fnom : MOT, Fprenom : MOT, n : N).

En supposant que les rôles des parents du mari soient codés 3 (père) et 4 (mère), la relation PMARI est le résultat de la requête SQL :

```
SELECT   a.nom as Mnom, a.prenom as Mprenom,
         b.nom as Fnom, b.prenom as Fprenom, a.n
FROM     MROLE a, MROLE b
WHERE    (a.n = b.n) AND (a.role = 3) AND (b.role = 4) ;
```

La relation PFEMME est générée selon le même principe. Pour les couples de chacune de ces relations portant un patronyme "TRA", l'acte de mariage sera recherché au moyen d'une différence ensembliste exprimée sur la clé sémantique *(Mnom, Mprenom, Fnom, Fprenom)*, soit respectivement (PMARI - COUPLE) et (PFEMME - COUPLE).

5.2.3. La filiation entre mariages des parents et mariages des enfants

Le calcul relationnel permet de représenter les liens de filiation entre les individus cités dans les actes. Les opérations proposées s'appliquent seulement si la clé sémantique *(Mnom, Mprenom, Fnom, Fprenom)* identifie un couple et un seul, ce qui peut conduire à isoler les cas pour lesquels des ambiguïtés demeurent.

La jointure des relations COUPLE et PMARI sur la clé *(Mnom, Mprenom, Fnom, Fprenom)* rend une relation qui représente la filiation du mari, soit :

FMARI (Mnom : MOT, Mprenom : MOT, Fnom : MOT, Fprenom : MOT, parent : N, enfant : N).

L'attribut *parent* contient le numéro de l'acte de mariage des parents, extrait de la relation COUPLE. L'attribut *enfant* contient le numéro de l'acte de mariage d'un enfant, extrait de la relation PMARI. Le résultat est obtenu en exécutant la requête SQL suivante :

```
SELECT   a.Mnom, a.Mprenom, a.Fnom, a.Fprenom,
         a.n as parent, b.n as enfant
FROM     COUPLE a, PMARI b
WHERE    (a.Mnom = b.Mnom) AND (a.Mprenom = b.Mprenom)
          AND (a. Fnom = b.Fnom) AND (a.Fprenom = b.Fprenom).
```

La relation qui représente la filiation de la femme, FFEMME, est construite de la même façon au moyen des relations opérandes COUPLE et PFEMME.

Ces relations de filiation représentent, en ascendances masculine et féminine, les graphes (ou relations binaires) qui relient les mariages d'une génération à ceux de la génération précédente.

Dans l'hypothèse d'un relevé exhaustif, les relations finales contiennent pour chaque couple parent autant de tuples qu'il y a d'enfants mariés.

L'**intersection** des relations FMARI et FFEMME désigne les mariages entre TRA. L'**union** de ces mêmes relations fournit l'ensemble des TRA pour lesquels l'acte de mariage des parents est présent.

5.2.4. Représentation des filiations sur les éléments de l'ensemble des TRA mariés

Les tuples éléments des relations précédentes représentent des familles. Le couple d'attributs *(parent, enfant)* traduit une relation binaire entre familles : les valeurs de ces attributs identifient de façon non ambiguë les actes de mariage des parents et des enfants. Cependant, pour la première génération du début du XIX^{ème} siècle, il n'est pas possible de faire référence au mariage des parents : par convention, l'attribut *parent* prend alors une valeur négative.

L'enquête ne retient que la descendance en lignée masculine (le cas des mères célibataires pouvant être traité à part). Ces lignées sont alors représentables dans la relation :

MTRA (Mnom : MOT, Mprenom : MOT, Fnom : MOT, Fprenom : MOT, parent : N, enfant : N).

Cette relation représente aussi un graphe sans circuit, un arc ayant comme point de départ la référence de l'acte de mariage d'un enfant et comme point d'arrivée la référence de l'acte de mariage du père légal.

En tout point part un arc et un seul. Les plus longs chemins ont pour longueur le nombre maximum de générations depuis le début de l'enquête. Tous les chemins ont comme point d'arrivée la référence à un mariage de la première génération du XIX^{ème} siècle. La recherche peut donc se faire en activant des règles de parcours récursives, le nombre d'itérations étant borné par la longueur du plus long chemin.

Le langage SQL permet seulement de parcourir un arc à la fois. Des extensions procédurales existent mais demeurent propriétaires. Les solutions sont éventuellement à rechercher autour des mécanismes de gestion des règles d'intégrité, qui font l'objet d'une normalisation.

5.2.5. Représentation des filiations des TRA mariés à partir des actes de naissance

Comme pour les actes de mariages, l'identification des couples parents est calculée dans la relation :

PARENT (Mnom : MOT, Mprenom : MOT, Fnom : MOT, Fprenom : MOT, lieu : N, date : N, n : N).

Cette relation est produite par une autojointure et sur la relation NPARENT, au moyen d'une requête SQL :

```
SELECT    a.nom as Mnom, a.prenom as Mprenom,
          b.nom as Fnom, b.prenom as Fprenom,
          a.lieu, a.date, a.n
FROM      NPARENT a, NPARENT b
WHERE     (a.n = b.n) AND (a.role = 1) AND (b.role = 2) ;
```

La référence au mariage d'un enfant est conservée dans la relation NMARIAGE. Ces informations sont suffisantes pour construire un ensemble de lignées partielles, représentable dans la relation :

NTRA (Mnom : MOT, Mprenom : MOT, Fnom : MOT, Fprenom : MOT, parent : N, enfant : N).

L'attribut *parent* fait référence à l'acte où figure la mention marginale du mariage de l'enfant TRA. La clé sémantique *(Mnom, Mprenom, Fnom, Fprenom)* reçoit les valeurs extraites de la relation NMARIAGE pour l'acte de naissance désigné par la clé parent. L'attribut *enfant* désigne un acte dans lequel *(Mnom, Mprenom)* représente le père et *(Fnom, Fprenom)* la mère de l'enfant TRA.

A une valeur de l'attribut *parent* (laquelle fait référence à l'acte de naissance du père) correspond un ensemble de valeurs de l'attribut *enfant* (normalement le nombre d'enfants du couple). Quand il y a plusieurs mariages, la valeur de la clé sémantique permet de distinguer les couples parents. La relation NTRA représente un graphe sans circuit, un arc reliant l'acte de naissance d'un enfant à celui de son père légal.

L'initialisation de la relation NTRA peut être faite au moyen de la requête SQL suivante :

```
SELECT    b.Mnom, b.Mprenom, b.Fnom, b.Fprenom,
          c.n as parent, a.n as enfant,
FROM      PARENT a, NMARIAGE b, PARENT c
WHERE     (a.n = b.n) AND (b.Mnom = c.Mnom) AND (b.Mprenom = c.Mprenom)
          AND (b.Fnom = c.Fnom) AND (b.Fprenom = c.Fprenom) ;
```

Les éléments pour lesquels l'acte de naissance des parents est inconnu sont ajoutés, l'attribut *parent* recevant une valeur négative.

5.2.6. Représentation des généalogies

Les relations MTRA et NTRA sont des graphes dans lesquels les sommets sont des individus, repérés par une clé sémantique. Les arcs sont générés en recherchant l'identifiant d'individu associé à la valeur de l'attribut *parent*. Cette recherche s'effectue dans FMARI pour les mariages, dans NPARENT pour les naissances. L'identifiant doit être unique. Cette

contrainte traduit la propriété usuellement appelée **intégrité de référence**. Avec le formalisme de l'algèbre relationnelle, on vérifiera que l'attribut *parent* est clé étrangère sur MTRA et NTRA, et clé primaire sur FMARI ou NPARENT. L'existence de ces propriétés est fondamentale, les résultats étant indéterminés en cas de violation.

L'objectif recherché est le raccordement des arborescences. Les clés instrumentales utilisées renvoient aux actes. Il faut définir une autre clé instrumentale qui désigne les individus, jusqu'alors repérés par une clé sémantique plus parlante pour le généalogiste.

Nous faisons l'hypothèse que les clés sémantiques sont aussi des clés primaires : cela revient pratiquement à isoler les ambiguïtés pour les traiter à part. La construction de la clé instrumentale s'effectue en énumérant les valeurs des quadruplets *(Mnom, Mprenom, Fnom, Fprenom)* et en leur affectant un numéro d'ordre, ce qui traduit en fait la notation indiciaire usuelle. Cette action fait appel à une fonction de marquage [9]. Le résultat est la relation :

DICO (Mnom : MOT, Mprenom : MOT, Fnom : MOT, Fprenom : MOT, i : N).

Le quadruplet *(Mnom, Mprenom, Fnom, Fprenom)* est aussi une clé primaire avec les conventions adoptées.

L'échantillon TRA est obtenu en enchaînant plusieurs jointures qui recherchent d'abord les clés sémantiques, puis les clés instrumentales désignant les individus. Les clés sémantiques peuvent être conservées pour faciliter la tâche du généalogiste mais sont théoriquement inutiles. Elles peuvent être supprimées par projection et régénérées en partie par un mécanisme de vues. Dans toutes ces opérations, il est indispensable de contrôler les cardinalités des résultats intermédiaires, une erreur traduisant généralement une violation de l'intégrité de référence. Le résultat final est une relation de schéma : GTRA (iparent : N, ienfant : N).

La recherche des composantes connexes du graphe à partir des données actuellement disponibles va mettre en évidence l'impossibilité de raccorder les lignées pour les enfants nés de mariages célébrés au début du XX^{ème} siècle.

---

[9] Ces méthodes ont été utilisées à l'INRA pour régler des problèmes d'harmonisation de nomenclatures sur des projets de recherches relatifs à la consommation alimentaire.

6. CALCUL ET EXTRACTION DE L'ECHANTILLON

La construction du graphe décrit dans le chapitre précédent impose d'identifier tous les individus susceptibles d'appartenir à l'échantillon, soit l'ensemble des TRA, puis d'ajouter les arcs qui traduisent la relation "x descend de y".

Le modèle proposé crée les sommets et les arcs en réalisant des tests d'équivalence de clés sémantiques, construites avec des noms propres nés d'une tradition orale. Sur le plan logique, l'égalité est vérifiée si et seulement si les tableaux de caractères associés aux attributs des clés sont rigoureusement identiques, le test étant réalisé caractère par caractère. Cette notion d'égalité est inadaptée au contexte des données patronymiques. Elle est peu efficace pour construire les arcs et aboutit de surcroît à créer plusieurs sommets pour un même individu.

6.1. La syntaxe des patronymes et prénoms

Les patronymes apparaissent comme des noms propres ou des groupes nominaux ; les particules demeurent ou se concatènent aux noms propres dans les générations suivantes et au hasard des migrations géographiques. Les composants terminaux des groupes nominaux complexes sont souvent supprimés. Sur ces problèmes de formes syntaxiques existent des conventions formalisées dans des normes ou des standards de fait. La construction de la primitive d'extraction des noms propres nécessite de connaître une liste exhaustive de règles logiques, à traduire sur les structures de données retenues.

Un patronyme est représenté dans un système informatique par une suite contiguë de caractères choisis parmi les 26 lettres de l'alphabet latin, le caractère d'espacement et accessoirement d'autres caractères comme le tiret et l'apostrophe. Les lettres accentuées et la distinction minuscule-majuscule ont été perdues.

Le résultat de cette étape préliminaire est une grammaire formelle pour les patronymes, à traduire dans un langage de programmation.

Une description des formes syntaxiques est également nécessaire pour les prénoms. Dans les fichiers usuels, les lettres isolées "J" et "M" remplacent respectivement les prénoms "Jean" et "Marie". Un choix pragmatique, à valider, consiste à retenir un prénom double lorsque le premier prénom est "Jean" ou "Marie" et un second prénom existe. Un ensemble de règles minimales, ne serait-ce que l'appartenance des caractères à l'alphabet, doit donc être construit pour définir une grammaire formelle relative aux prénoms.

6.2. Les dictionnaires des patronymes et prénoms

L'enquête référence des ensembles finis de noms et prénoms, définis par une syntaxe et répartis dans des relations. Les noms et prénoms reconnus sont rangés dans des dictionnaires, utilisables pour les contrôles puis pour l'analyse des équivalences des formes écrites. Un dictionnaire simple est représentable dans la relation : DICO (nom : MOT, s : N).

L'attribut *s* dénombre les occurrences associées à chaque valeur de l'attribut *nom*, clé primaire de la relation. Le calcul s'effectue par union de toutes les listes de noms d'une classe donnée (patronymes ou prénoms) avec agrégation (dénombrement ou somme selon le protocole opératoire). Des stratégies de contrôle pourront être développées au moyen de différences ensemblistes sur des dictionnaires plus détaillés qui prendraient en compte les rôles des individus, une même personne étant souvent citée dans plusieurs documents.

6.3. La numérisation des lieux

Un lieu est identifié dans un document d'état-civil par un nom de commune dans un département. Tout lieu peut être désigné de façon non ambiguë par le code INSEE de la commune [10]. Le dictionnaire des communes est représenté dans la relation : COMMUNE (nom : MOT, code : N).

L'attribut nom contient toutes les variantes orthographiques rencontrées pour l'écriture des noms de communes.

Une différence ensembliste rend la liste des lieux cités dans les documents mais non répertoriés dans la relation COMMUNE. Après correction ou ajout de tuples dans le dictionnaire jusqu'à ce que la différence ensembliste soit vide, une opération de jointure sur les attributs contenant les noms de lieux en clair introduit le code numérique du lieu dans les relations d'archivage[11].

6.4. Les classes d'équivalence des noms propres

Un individu hérite du patronyme de son père. Cette règle s'applique de génération en génération. Or, cette entité héritée, le patronyme, se moule difficilement à une approche formelle. La traduction écrite, syntaxiquement normalisée, subit des variations liées aux

---

[10] Pour la Corse, les codes "2A" et "2B" sont remplacés par "20", pour définir un code commune numérique.
[11] Cette méthode est déjà utilisée à l'INRA dans le cadre du programme PAGI (Base de données de l'état civil ancien).

différentes façons de transcrire les phonèmes de la langue orale. Nous supposons que les variations sur les phonèmes sont négligeables sur la période d'observation. Dans un premier temps, nous ignorons ces variations dans l'espace géographique : la notion de région peut cependant être partiellement prise en compte pour des cas résiduels, en utilisant l'attribut *lieu* des relations d'archivage des documents. Le problème à résoudre est de construire un opérateur efficace pour tester l'équivalence de deux patronymes. La question se pose mais avec des variations beaucoup plus restreintes pour les prénoms.

La première étape consiste à énumérer les règles usuelles d'écriture des noms de la langue française : transcription des phonèmes, doublement de consonnes, évolution des voyelles accentuées, terminaison des mots, etc... . Un manuel de grammaire de l'enseignement primaire suffit. Le choix des règles à retenir mobilise des compétences d'expert. La traduction de ces règles utilise des techniques de compilation. La disponibilité des données fait qu'il est possible d'expérimenter plusieurs stratégies, pour aboutir à un compromis entre le scénario silencieux (qui produit peu de liens) et le scénario bruyant (qui associe plusieurs pères à un même individu).

Le seconde étape utilise l'opérateur ainsi validé pour comparer deux à deux tous les mots du dictionnaire des patronymes. Cet opérateur traduit une relation d'équivalence. Chaque classe d'équivalence contient les variantes orthographiques d'un même patronyme. Le représentant d'une classe d'équivalence, dit "nom normalisé", pourrait être la forme la plus fréquemment citée.

Les patronymes sont alors numérisables, chaque classe d'équivalence étant repérée par un code numérique instrumental : le nom de la classe est restitué par un mécanisme de vue, ou table virtuelle.

Cette méthode génère au moyen d'un système de règles une table des noms normalisés, qui contient aussi les patronymes des conjoints des TRA. Elle s'applique également, avec un système de régles réduit, à l'ensemble des prénoms. Les noms et prénoms normalisés sont introduits dans des relations virtuelles sur lesquelles sont recherchés les liens entre individus, comme décrit dans le chapitre précédent.

7. L'ARCHITECTURE DE LA BASE DE DONNEES DEMOGRAPHIQUES

Les observations démographiques concernent des variables liées aux familles issues de l'échantillon initial des 3 000 familles de la première génération. Elles sont hiérarchisées : familles et individus. Elles sont réparties dans plusieurs documents et codées dans des segments logiques, conservés dans leur intégralité dans une base d'archivage.

On dénombre 11 000 mariages TRA entre 1803 et 1832, dont 6 200 fils et 4 900 filles. L'échantillon initial des familles est à extraire de la population des mariages des fils TRA, soit globalement un mariage sur deux. Cet échantillon croît de génération en génération par ajout des familles issues des mariages des fils. Les filles n'interviennent pas dans cette construction dynamique : elles sont seulement présentes comme enfant dans une famille de l'échantillon.

7.1. Gestion de la relation d'appartenance à l'échantillon

Les familles sont normalement désignées par les noms et prénoms usuels des époux, au moyen des attributs *Mnom, Mprenom, Fnom, Fprenom*. L'appartenance à l'échantillon est codée dans l'attribut numérique *drapeau*, toute valeur négative (-1) exprimant une indétermination. Le numéro de génération est conservé dans l'attribut *g*, la première génération étant choisie sur la période 1803 - 1832. Le schéma de relation est finalement le suivant :

ECHANTILLON (Mnom : MOT, Mprenom : MOT, Fnom : MOT, Fprenom : MOT, g : N, drapeau : N).

La clé primaire *(Mnom, Mprenom, Fnom, Fprenom)* désigne logiquement l'ensemble des familles civilement identifiées par les noms et prénoms associés à la valeur de la clé.

La population de départ est l'ensemble des familles TRA constituées lors d'un mariage de première génération (g = 1). Une condition nécessaire et suffisante pour qu'un mariage soit de première génération est qu'il soit célébré sur la période 1803 - 1832 et que le mariage du père n'ait pas été célébré sur cette même période. Une telle propriété s'exprime simplement au moyen du langage SQL. La relation ECHANTILLON est initialisée en tirant 3 000 familles dans la population de départ et en affectant la valeur 1 dans l'attribut *drapeau*. Les familles non retenues mais appartenant à la population de départ sont ajoutées à la relation échantillon avec une valeur nulle pour l'attribut *drapeau*. A ce stade, l'attribut *drapeau* est défini et a trois valeurs possibles : - 1 pour une indétermination, 0 pour la non appartenance à l'échantillon et 1 pour l'appartenance.

La construction de l'échantillon se poursuit par ajout de la génération 2. Une famille TRA appartient à la génération 2 (g = 2) si la famille du père est de génération 1. Pour tout mariage d'un fils TRA, on recherche dans l'ensemble ECHANTILLON la référence des parents

avec la restriction (g = 1). Si les parents sont présents, la famille issue du mariage est ajoutée, l'attribut *g* recevant la valeur 2 et l'attribut *drapeau* héritant de la valeur de l'attribut *drapeau* des parents. Le processus se poursuit, une famille de génération k étant ajoutée si et seulement si elle est créée par un fils TRA d'une famille de génération (k - 1), appartenant à l'ensemble ECHANTILLON : elle hérite de la valeur de l'attribut *drapeau* de la génération précédente.

La relation ECHANTILLON joue un rôle fondamental dans la conduite de cette enquête. Elle permet, pour tout identifiant de famille *(nom1, prenom1, nom2, prenom2)* extrait d'un document quelconque, de tester l'appartenance à une lignée issue des 3 000 familles initiales. Il suffit d'exécuter la requête SQL :

```
SELECT    g, drapeau
FROM      ECHANTILLON
WHERE     (Mnom = nom1) AND (Mprenom = prenom1) AND (Fnom = nom2)
          AND (Fprenom = prenom2) ;
```

Si le résultat est vide, il y a indétermination. Dans le cas contraire, la valeur non négative de l'attribut *drapeau* du tuple rendu indique si la famille est présumée appartenir (drapeau = 1) ou ne pas appartenir (drapeau = 0) à l'échantillon. Cette présomption résulte de l'existence possible de familles hors échantillon qui seraient identifiées par les mêmes noms et prénoms. La puissance du test est à évaluer.

7.2. Organisation des espaces d'observations

L'enquête est définie sur un échantillon de familles, la famille désignant un couple civil et ses enfants. L'espace initial est donc celui des familles de l'échantillon. Une personne crée plusieurs familles si elle se marie plusieurs fois.

Les observations élémentaires concernent principalement les individus cités dans les actes d'état-civil : un second espace est celui des individus. Les espaces des familles et des individus sont représentables dans une modélisation hiérarchique. L'individu hérite des variables de la famille. Inversement, l'agrégation de variables relatives aux individus génère des variables descriptives des familles. Toute famille est identifiable et l'individu est lui-même identifiable dans sa famille.

Ce modèle théorique, construit sur les entités famille et individu, se heurte à des difficultés pratiques liées à la source d'informations. Les renseignements relatifs à un même individu sont dispersés dans les actes : une personne civile est d'abord sujet de son acte de naissance, puis de mariage ; elle apparaît ensuite comme parent pour les naissances et les

mariages de ses enfants; elle est enfin sujet de son acte de décès. Les observations associées fournissent une chronologie, de périodicité irrégulière, entre la naissance et la mort. A chaque citation d'un individu dans un acte est associé un segment logique extrait d'un enregistrement : les valeurs des variables sont codées dans un ensemble de segments, qui peut regrouper plusieurs dizaines d'éléments.

Le calcul d'une valeur d'une variable élémentaire observée pour un individu à une période donnée passe par le traitement de cet ensemble de segments. Les règles de contrôle de cohérence sont à décrire, une valeur étant réputée fiable si elle est homogène dans tous les segments. L'énoncé de ces règles, et le choix éventuel des seuils dans des situations d'incertitude, relèvent de la responsabilité de l'équipe scientifique de pilotage. Nous soulignons le fait que la redondance des informations doit être pleinement exploitée, dans la mesure où elle compense les carences liées à la nature des documents de base.

7.3. Choix et extraction des variables

Dans le système d'archivage proposé, les identifiants des familles sont des clés d'accès sur des ensembles de segments d'informations, qui contiennent les valeurs des variables élémentaires. Le format de ces segments n'est que partiellement connu : seules les informations d'identification ont été extraites. Certaines zones ont une signification constante: libellés de métiers, dates, lieux, codes explicites (S pour signature, L pour légitime, N pour naturel, N pour naissance...) . Des champs sont systématiquement remplis alors que d'autres sont généralement vides. Au prix de tâtonnements et avec l'aide d'un moteur relationnel efficace se profilent progressivement des formats hypothétiques : un champ de 4 chiffres qui commence par le chiffre 1 et est suivi des chiffres 8 ou 9 contient une année, une position qui ne contient que les lettres S ou N code la signature... .

Nous avons développé quelques utilitaires simples de décryptage et d'édition autour d'un SGBDR, pour tester des hypothèses sur le format, repérer des anomalies apparentes et réaliser des dénombrements.

La démarche suivie revient à décrire a posteriori et au vu des fichiers les variables élémentaires, les domaines de définition associés et les formats des enregistrements logiques. Elle doit aboutir à l'énoncé d'un ensemble de règles pour les contrôles de cohérence et au choix définitif d'une liste de variables.

Si les bases d'archivage conservent l'intégralité de l'information non contrôlée et non structurée, la base démographique contient une information structurée et validée, affectée dans des variables définies sur un espace d'observation. Les données de la base démographique sont organisées pour pouvoir être traitées par les logiciels statistiques les plus communément utilisés. La production de cette base nécessite de réaliser la traduction informatique de règles syntaxiques et propriétés sémantiques mises en évidence au cours d'une phase d'analyse. Elle passe par l'écriture et la validation d'une série de primitives sur les tableaux de caractères issus des segments logiques. A tout moment, il doit être possible d'intégrer de nouvelles variables extraites des bases d'archivage et qui n'auraient pas été retenues pour des raisons d'opportunité ou de coût dans le choix initial.

7.4. Principes pour la structuration de la base démographique

L'objectif principal de ce rapport est la formalisation d'une méthode opérationnelle pour établir un état des lieux rigoureux, au stade actuel de l'enquête. Le contenu de la base démographique n'étant pas encore arrêté, il serait prématuré d'établir des schémas de relations définitifs. Nous pouvons cependant proposer un modèle générique et nous prononcer sur sa faisabilité.

7.4.1. La numérisation des variables et des identifiants de personnes

La base démographique est intégralement numérisée, les données textuelles étant remplacées par des valeurs numériques. Les correspondances sont gérées dans des dictionnaires. Un mécanisme de vue permet de restituer les valeurs textuelles pour des besoins applicatifs. Cette numérisation, outre qu'elle réduit les volumes physiques, restreint les domaines de définition aux ensembles de nombres, et permet l'utilisation des opérateurs arithmétiques usuels dans des expressions parenthèsées. Elle s'applique aussi aux clés sémantiques textuelles utilisées pour identifier les individus dans la base d'archivage des documents et offre ainsi un moyen efficace pour assurer la **confidentialité** des données. Les liens entre l'individu statistique de la base démographique et la personne citée dans un acte sont conservés à l'extérieur, pour les seuls besoins de production de la base et de maintenance. Pratiquement, ces liens sont gérés dans la relation :

ITRA (Eprenom : MOT, Mnom : MOT, Mprenom : MOT, Fnom : MOT, Fprenom : MOT, g : N, nf : N, np : N).

Le numéro de génération *g*, hérité, est ajouté pour réduire les possibilités d'ambiguités tout en suivant les générations successives. Les attributs *nf* et *np* sont des codes instrumentaux qui désignent des numéros de famille et des numéros de personne dans la famille attribués par un algorithme. La clé primaire *(g, nf, np)* va permettre la sélection dans d'autres relations de toutes les informations relatives aux individus.

7.4.2. La représentation des données démographiques

L'ensemble des données démographiques, pour tous les individus et sur toutes les périodes, est représentable dans la relation :

INDIVIDU (g : N, nf : N, np : N, role : N, nv : N, t : N, x : N)

L'individu, identifiable au moyen de la clé étrangère *(g, nf, np)*, intervient avec un statut codé dans l'attribut *role*. Les variables sont identifiées par l'attribut numérique *nv*, lequel est clé étrangère dans un dictionnaire des variables qui contient les noms symboliques et une documentation. La dimension temporelle est prise en compte au moyen de l'attribut *t* qui situe l'information dans le temps. La valeur de la variable est rangée dans l'attribut *x*. Le couple *(nv, x)* est souvent clé étrangère dans un dictionnaire de définition de modalités (lieux, métiers...).

Nous retenons comme clé primaire le quintuplet *(nf, np, role, nv, t)*. Avec cette convention, on peut séparer les informations relatives à un individu sujet d'un acte de décès le jour du mariage d'un enfant. En revanche, on ne sépare pas les informations extraites de deux actes de mariage pour un père qui marierait deux enfants le même jour. Cela signifie que la cohérence des informations est à vérifier en amont : la structure adoptée n'autorise pas la présence d'informations différentes en provenance de deux sources distinctes.

Ce modèle générique présente l'avantage d'être dynamique, du fait qu'il est toujours possible d'ajouter ou supprimer variables et observations.

7.4.3. La gestion de filiations

Tout individu de la génération *g* est lié à la famille issue de son père légal, unique, de génération *g-1*. Ce fait généalogique se traduit dans la relation :
PAPA (g: N, nf : N, np : N, famille : N). L'attribut *famille* est le numéro de famille de la génération *g-1*.

7.4.4. Les liens avec les bases archives

Les relations de la base archive sont des relations associées à des sources de documents, identifiables par un code numérique. Le lien avec la base démographique est géré dans la relation :

SOURCE (g : N, nf : N, np : N, role : N, s : N).

L'attribut *s* va désigner une relation de la base archive. Cette modélisation offre des possibilités de paramétrage dans des requêtes SQL dynamiques, "encapsulées" dans des programmes, pour construire des interfaces adaptées aux besoins scientifiques.

7.5. L'analyse de complexité

Le plus grand ensemble de cette base est la relation INDIVIDU, qui regroupe toutes les personnes liées à l'échantillon, observées dans tous les actes de l'état-civil les concernant. Avec un nombre de variables par individu cité dans un acte de l'ordre de la dizaine, la cardinalité de cet ensemble n'atteint pas les huit millions ($\approx 2^{23}$). L'ordonnancement d'un tel ensemble est rapide, sous réserve évidemment que le moteur relationnel soit capable de choisir un algorithme efficace dans ce contexte [12]. Quant au volume, il serait de 160 mégaoctets pour un fichier qui utiliserait les types numériques usuels. Nous en concluons qu'une station de travail et un moteur relationnel efficace sur de grands ensembles sont suffisants pour construire le noyau de la base démographique.

7.6. L'accès à la base démographique

Le modèle logique proposé gère l'intégralité de l'information dans des structures algébriques simples. La restitution de l'information pour une utilisation en sciences sociales met en oeuvre des opérations de calcul ensembliste pour réaffecter aux codes instrumentaux des données textuelles. Ces opérations s'expriment au moyen du langage SQL mais demandent une connaissance minimale des règles de calcul logique et une certaine pratique. Les données élémentaires extraites sont ensuite regroupées pour générer par exemple une fiche de famille, au format d'édition POSTSCRIPT, visualisée sur un écran graphique ou imprimée.

La réalisation de telles interfaces pour des disciplines scientifiques est une tâche complexe qui se situe en aval de cette première grande étape de production des relations fondamentales de la base démographique.

---

[12] La fonction de complexité associée à l'algorithme de tri doit impérativement être en *nlog(n)*.

8. LES MOYENS INFORMATIQUES

"La mise à disposition à temps partiel d'un responsable informatique de haut niveau" et la constitution "d'une force de travail qui aurait à construire concrètement les outils informatiques nécessaires" [13], comme proposé dans le rapport Villac, **ne sont pas suffisantes pour produire la base démographique initiale**. Une partie importante des méthodologies et conceptualisations nécessaires relève des sciences de l'information : la réalisation est un problème original d'ingénierie informatique, pour lequel il ne suffit pas "de fixer les grandes lignes de la structure matérielle et logicielle à mettre en place, et de contrôler sa réalisation" [14]. S'il y a place pour des prestations d'analyse-programmation classiques et des vacations d'étudiants de niveau mastère pour construire des composants, **il est indispensable de mobiliser sur une période continue et suffisamment longue des qualifications de niveau "Ingénieur de Recherche" en informatique.**

8.1. Moyens matériels et logiciels

Le poste de travail du développeur est une station de travail UNIX (64 Mo de mémoire, 2 Go de disque, écran 19") avec une interface XWINDOW MOTIF. Les opérations de saisie sont réalisées sur microordinateurs avec des logiciels spécialisés, les fichiers étant ensuite rapatriés sur station UNIX. L'accès aux stations de développement pourra se faire au moyen d'une émulation XWINDOW : pour un travail continu avec une utilisation intensive des interfaces graphiques XWINDOW, un terminal X est recommandé [15].

Le SGBDR retenu doit impérativement disposer d'un optimiseur statistique efficace, pour choisir les stratégies algorithmiques et de stockage adaptées à la taille des ensembles manipulés. Le produit INGRES, utilisé à l'INRA, répond convenablement à ce besoin. Il offre en plus la gestion des types abstraits de données et un langage graphique, WINDOWS4GL, qui valorise pleinement les possibilités de XWINDOW MOTIF. Ce langage sera utilisé pour construire des interfaces entre les versions successives des bases de travail et l'expert chargé de contrôler la cohérence et rechercher les liens.

L'utilisation d'approches orientées objet n'est pas prévue à l'heure actuelle sans être exclue pour autant. La solution proposée s'appuie en fait sur des concepts stabilisés dont beaucoup font l'objet de normes ou standards.

---

[13] Rapport Villac, p. 28.
[14] Rapport Villac, p. 28.
[15] Les langages graphiques à la périphérie des SGBD génèrent la plupart du temps un nombre important de fenêtres et widegts.

8.2. Les ressources humaines en informatique

Cette évaluation ne concerne pas les tâches de saisie, les données étant récupérées dans des fichiers. Elle exclut également les opérations lourdes de correction. Le responsable informatique à temps partiel est chargé d'initialiser le processus puis d'en suivre le déroulement et de valider les différents modules. Son activité est importante au début du projet puis devrait se réduire progressivement.

L'équipe de développement comportera au moins un ingénieur de recherche à mi-temps, assisté d'un analyste-programmeur. Des vacations d'étudiants et des contributions d'ingénieurs apporteront une force de travail supplémentaire pour construire et valider des composants.

8.3. L'ordonnancement des tâches

En fonction de ce qui précède, on peut distinguer trois grandes étapes :

étape 1 : construction de la base archive ;

étape 2 : modélisation logique et construction de la base démographique ;

étape 3 : réalisation d'interfaces pour l'accès aux données et maintenance des instruments.

En fait, l'objectif principal, à savoir "mettre une base de données à la disposition de la communauté scientifique", sera atteint dès la fin de l'étape 2. Toutefois, les utilisateurs potentiels seront en présence d'une représentation abstraite, sous forme d'un ensemble de relations, sans qu'ils puissent forcément voir les objets qu'ils manipulent habituellement au sein de leur discipline scientifique : tableaux statistiques, fiches de famille, arbres généalogiques, etc... .

8.3.1. Construction de la base archive

La base archive permet d'identifier toutes les entités présentes dans tous les actes et de relier des entités entre elles. Elle conserve l'intégralité des informations, tout acte pouvant toujours être reconstitué. Seules les informations d'identification des actes et des individus sont restituées sous une forme atomique. Les variables descriptives demeurent codées dans des segments logiques, sous une forme qui pourrait être qualifiée de "moléculaire". La restitution des valeurs élémentaires des variables descriptives relève de la seconde étape.

La masse des informations disponibles est largement suffisante pour démarrer immédiatement la construction de la base archive : une première expérimentation en vraie grandeur a précédé l'écriture de ce rapport.

Cette première étape enchaîne trois tâches principales :

tâche 1 : reconnaissance syntaxique des enregistrements logiques et apuration pour toutes les sources de données ;

tâche 2 : construction du moteur pour la génération automatique des liens de filiation ;

tâche 3 : développement des interfaces pour l'expert.

Les deux premières tâches peuvent être réalisées parallèlement et l'on peut considérer qu'elles sont initialisées pour les actes de mariage. Elles s'appuient sur la logique ensembliste (calcul relationnel) et les grammaires formelles, et utilisent le langage SQL et des techniques usuelles de compilation.

La dernière tâche suppose que les deux premières sont suffisamment avancées. Elle utilise les interfaces graphiques XWINDOW MOTIF pour dialoguer avec les experts de la discipline. En effet, si la machine propose des automatismes, la validation relève de la responsabilité de l'expert. Les interfaces fournissent des outils graphiques pour assister l'expert dans la résolution des ambiguïtés susceptibles d'intervenir dans une filiation. Elles proposent normalement des tests activables pour repérer des contradictions qui résulteraient de choix erronés ou d'ambiguités non résolues. La conception de telles interfaces est complexe et mobilise des compétences spécialisées. Les compétences requises sont celles plus couramment utilisées dans la conception des postes de supervision d'installations complexes ou de pilotage.

En raisonnant en temps-homme moyen de la force de travail nécessaire (ingénieur de recherche + analystes-programmeurs), une évaluation empirique donne 9 mois-homme pour chacune des deux premières tâches et 12 mois-homme pour la dernière, soit 30 mois-homme au total.

8.3.2. Construction de la base démographique

A l'issue de la première étape, la base archive va contenir un ensemble d'informations suffisant pour rechercher au moyen d'opérations ensemblistes, pour le XIX^{ème} siècle d'abord, les actes de naissance et décès et les actes de mariage manquants. Si le choix de l'échantillon de première génération est définitivement arrêté, cette recherche pourra être restreinte à l'échantillon généré, soit pratiquement un TRA sur deux. Pour le XX^{ème} siècle, il faudra attendre la disponibilité des actes de mariage. La base archive s'enrichira progressivement par

ajout de données et calcul de liens, au moyen des instruments de l'étape précédente pilotés par un expert.

Cette seconde étape est divisible en trois familles de tâches :

tâche 1 : extraction et construction des variables descriptives ;

tâche 2 : contrôle des domaines de définition et apuration ;

tâche 3 : numérisation et validation du modèle logique.

La première tâche va consister à extraire toutes les variables des segments et à en évaluer la qualité ; mais aussi à calculer des variables essentielles par regroupement de segments : rang d'un enfant, existence d'un métier permanent... . Elle utilise les acquis de la première étape et se situe dans son prolongement. L'information est stockée sous sa forme atomique, c'est-à-dire une valeur associée à une observation complètement localisée : un individu peut avoir deux métiers différents dans des actes proches dans le temps. Le principe retenu est de conserver toutes les valeurs atomiques.

La numérisation de la base démographique nécessite de construire des dictionnaires de définition, lesquels sont aussi des outils utilisables par l'expert pour des besoins de contrôle et d'harmonisation (choix de représentants des classes d'équivalence, valeurs jugées aberrantes...). Ces contrôles s'effectuent au moyen d'opérations ensemblistes. Les dictionnaires seront également utilisés dans les "vues" de la base démographique pour restituer les données dans l'environnement des thématiques scientifiques. La qualité de ces dictionnaires est donc essentielle et conditionne la qualité du résultat final.

La dernière tâche valide définitivement le modèle logique et charge les données. Toutes les opérations qui suivront vont dépendre de la qualité et de l'efficacité du modèle adopté.

Globalement, ces tâches nécessitent approximativement, pour le volet strictement informatique, environ deux années-homme.

8.3.3. Réalisation d'interfaces et maintenance des instruments

A ce stade, la machinerie devrait être opérationnelle, ses performances étant fortement liées au soin apporté dans la réalisation des étapes précédentes. L'arrivée massive des actes de naissance et décès pourra mettre en évidence des dysfonctionnements qui feront l'objet d'intervention de maintenance.

Le problème des interfaces avec les équipes de recherche se pose dans la mesure où l'accès aux données suppose la connaissance et la maîtrise du modèle logique. Quelques outils

seront donc à créer à la demande pour les plus fréquentes : édition POSTSCRIPT de fiches de familles ou d'arborescences, construction de tableaux statistiques, édition de fichiers pour les traitements... . En retenant un scénario minimal, la production d'interfaces étant réduite aux besoins élémentaires, cette dernière étape requiert au moins une année-homme.

CONCLUSION

Une condition nécessaire à l'aboutissement du projet est la reconnaissance de l'importance du volet "sciences de l'information". Il paraît illusoire de vouloir conduire ce projet à son terme en se contentant de recourir à des techniques mises en oeuvre par un "personnel intérimaire" : les spécialistes des sciences de l'information sont partie intégrante de l'équipe de recherche. Le risque est d'autant plus important que la popularité du langage SQL laisse souvent croire qu'il est possible de faire abstraction de la théorie mathématique sur laquelle repose le calcul relationnel. Les méthodes proposées dans ce rapport relèvent d'une démarche scientifique classique, et non de l'utilisation d'outils offerts "clés en main" sur le marché du logiciel. A ce titre, elles sont aussi des produits de la recherche qui font l'objet d'articles.

En conclusion, la production de la base de données démographiques liée à l'enquête TRA ne paraît pas soulever des difficultés théoriques insurmontables, mais elle n'aboutira pas si l'on se contente de moderniser des "méthodes artisanales", qui, dit le rapport Villac, "sont l'échelle classique des institutions de recherche (du moins dans les sciences sociales)" [16]. Il faudra mobiliser d'autres compétences pour mettre en place "des procédures plus proches de l'industrie" et produire un instrument informatique efficace au service des experts scientifiques des disciplines concernées.

---

16 Rapport Villac, p. 4.

TABLE DES MATIERES