LA RECHERCHE AUTOMATIQUE DANS NOS TABLES DE GÉNÉALOGIE

PAR A. ACCARY

1. INTRODUCTION

Serait-il possible, un jour, d'envisager de simplifier les opérations de recherche d'ancêtres ? A l'époque où la transmission du nom de famille est menacée par la législation même, on se demande où est l'intérêt de poser une telle question ?

Cependant, beaucoup de nos contemporains se jettent dans la quête de leurs ancêtres. Alors... goût éphémère, dernier sursaut d'une arrière-garde, désir de garder une trace des ancêtres, besoin sérieux de se replacer dans la société ancestrale qui y a t'il de sérieux dans nos démarches si fastidieuses de « rats d'archives anciennes » ?

Tout d'abord, celui qui a eu à souffrir de cette fièvre inquisitoriale, a du se poser très vite la question de la solidité des documents transmis par des internautes dans les sites de généalogie. Comment peut-on s'assurer de la validité d'une source ? Quel critère de qualité peut-on établir sur une banque de donnée, sur une liste d'ascendants... le nombre des individus ne fait pas la qualité d'un document.

Nous allons tenter de « caractériser » les tables de données relevées sur les registres paroissiaux et de l'état civil de deux communes : Mussy-sous-Dun et Chauffailles, en utilisant des méthodes de la logique. Le but de cet article est de rendre compte de la qualité des résultats que l'on peut obtenir des tables de Baptêmes et Sépultures. Nous espérons à terme disposer d'un outil de fiabilité connue, permettant de débuter une recherche dans quelques communes de notre région de l'Est-Brionnais.

2. LES TABLES D'ACTES

Notre travail a commencé par la photographie numérique des registres, beaucoup de fichiers numérisés sont maintenant disponibles pour une somme modique, sous forme de CD ROM[1]. L'étape suivante consista à trouver des lecteurs-traducteurs qui furent chargés d'établir la saisie, selon une grille précise, des informations (nom, prénom, date de l'évènement, ... , témoins à l'établissement de l'acte).

Nous avons utilisé une grille de saisie largement diffusée dans la région : Nimègue-version 2. Le travail fastidieux de traduction des actes est fait par ces nombreux collaborateurs qui se forment à la lecture (au déchiffrement) des actes manuscrits anciens. Les erreurs de transcription restent assez nombreuses, mais elles peuvent être corrigées « a postériori » par les internautes qui viennent consulter nos listes.

Afin de faciliter des opérations de correction ultérieures, nous avons choisi d'une part de donner l'accès à l'acte original, depuis son extraction d'une table et, d'autre part, de rajouter des colonnes de noms de familles, avec une forme orthographique « normalisée » unique et actuelle, qui jouxtent la colonne des noms fournis dans les

[1]chez : pour le 71 `http://suitegen.free.fr`, pour le dept. 69 `http://expgen.free.fr`

actes avec l'orthographe incertaine des prêtres des XVII et XVIIIè siècles[2]. Le recadrage des noms devrait faciliter la première opération de recherche.

Example 2.1. Toutes les orthographes des patronymes voisins : Bidaud, Bidot, Bideau, Bidaut, ... sont recodées dans une colonne supplémentaire sous une seule forme Bidaud, choisie pour être la plus représentée à notre époque.

Les tables ainsi saisies sont tout d'abord vérifiées et enregistrées dans une ***base de donnée relationnelle*** qui permet d'aller effectuer des opérations de tri sélectifs, selon des critères multiples choisis, reliés entre-eux par des ***opérateurs logiques***.

3. QUALITÉ DE LA BASE

Nous avons cherché à estimer la qualité de nos bases de données. Un premier travail consiste à comparer les nombres d'évènements relevés par nos traducteurs et ceux donnés par l'instituteur Jean Nesly dans son ouvrage désormais célèbre Base de Mussy sous Dun := `http://www.mussy-sous-dun.fr`.

3.1. **Comparaison de deux études.**

Au XVIIIè siècle, Jean Nesly a regroupé les naissances sur deux périodes de 50 ans, les valeurs indiquées par Nesly sont voisines de celles que nous avons relevé.

TAB. 1. Analyse comparative des naissances à Mussy sous Dun au XVIIIè siècle

Périodes	Jean Nesly	Nos relevés
1700-1750	1919	1980
1750-1800	2390	2333

Il apparaît une différence assez peu significative.

L'analyse comparative des nombres de décès, classés cette fois par tranches de dix années, donne encore des différences qui peuvent être plus significatives :

[2]Nous nous apercevrons par la suite qu'il faudrait faire le même rajout pour les prénoms, afin d'améliorer l'extraction des individus des tables de Baptêmes, Mariages et Sépultures.

TAB. 2. Analyse comparative des décès (tranches de dix ans)

Périodes	Jean Nesly	Nos relevés
1700-1709	316	341
1710-1719	142	151
1720-1729	123	183*
1730-1739	284	303
1740-1749	243	250
1750-1759	237	248
1760-1769	326	308
1770-1779	285	309
1780-1789	400	413
1790-1799	406	415

Les valeurs indiquées par Jean Nesly sont toujours légèrement inférieures aux notres, à l'exception des « années vingt » où il y a un décalage trop important !

4. RELATIONS D'IDENTIFICATION

Passons maintenant à la mise en place des relations d'identification des individus que l'on peut établir à partir des enregistrements de nos tables de données.

4.1. **Les relations d'identification.** Considérons un individu u, ***né*** et ***décédé*** à Mussy, caractérisé par une ligne de trente à quarante champs dans les tables de naissance et décès[3]. Afin de mettre en place une étude significative portant sur tous les individus, ayant pour but d'attribuer à l'un quelconque d'entre eux ses dates de naissance et de décès, il est nécessaire de d'utiliser les informations essentielles qui figurent dans nos tables : l'âge au décès, le sexe, les noms, prénoms.

Soit y un élément caractérisé dans la table de décès et x un autre dans celle des naissances.

Chercher à identifier un **individu** précis d'une manière **unique** revient donc à chercher les correspondances d'un certain nombre de champs de y avec ceux de x tels que la relation :

(4.1)
$$y = \{nom,\, prenom,\, sexe,\, date_{dc},\, age_{dc}\} \leftarrow x = \{nom,\, prenom,\, sexe,\, date_{naiss}\}$$

renvoit de la table des naissances dans celle des décès[4].

On cherchera à identifier un individu u à partir de l'égalité des nom, prénom, sexe, et de l'appartenance de l'âge calculé (e) à un voisinage de l'âge indiqué par le prêtre au moment du décès.

[3]Confère Nimègue-V2

[4]L'ensemble des individus (V) ayant vécu sur une période donnée à Mussy est l'intersection des deux tables des naissances et décès (B et S).

$$e \in \vartheta\left(age_{DC}\left(y\right)\right) \tag{4.2}$$

Cette relation (eq : 4.2) est calculée dans la requête SQL[5].
Une relation de correspondance devrait vérifier :

$$u\ identifie\ sss\ (y \leftarrow x)\ vrai$$

Example 4.1. *Parmi les éléments issus des relations $y \leftarrow x$ nous obtenons un y issu d'une identification avec les relations d'égalité des nom prenom et sexe, le calcul d'âge est vrai à 5% près :*

$$y = \{Auberger,\ Benoite,\ F,\ 19,\ 18/6/1771\} \leftarrow x = \{Auberger,\ Benoite,\ F,\ 7/10/1751\}$$
$$avec\, e = \frac{(date_{dc}(y) - date_{naiss}(y)) - age_{dc}}{age_{dc}} < a\,;\ a = 5\%$$

Cette relation implique donc que

$$u = Auberger\ Benoite,\ 1751 - 1771$$

Nota bene : La table des naissances peut encore donner les parents et on peut poursuivre l'identification des parents.

Cependant, certaines difficultés vont immédiatement apparaître :

(1) Les prénoms ne sont pas toujours orthographiés de la même manière dans deux tables d'actes ($\mathcal{B}$ et $\mathcal{S}$),
(2) Les prénoms peuvent être différents dans les deux actes (***Thomas*** Auberger peut devenir ***Claude*** Auberger mort à 50 ans, le calcul des dates donne un seul individu de sexe masculin nommé Auberger, décédé à 50 ans),
(3) Les âges peuvent être donnés, au moment du décès, de façon approximative (Jean Martin décédé à 50 ans pouvait en fait avoir 48 comme 52 ans !)[6]

4.2. **Les sous-ensembles.** Le classement que nous avons choisi d'appliquer repose sur le **regroupement** des individus de la table des décès en ***sous-ensembles de noms*** obtenus à partir des critères d'***analyse des âges*** de ces individus, au moment de leur décès[7].

Proposition 4.2. *Si l'âge d'un individu identifié y est connu, on le range dans S_k*
Si l'âge n'est pas connu, on le range dans S_u.

4.3. **Présentation de la démarche d'identification.** La démarche suivie est la suivante :

(1) On détermine dans la table des décès le nombre N d'individus dont l'âge est noté par le prêtre au moment du décès, survenu entre 1750 et 1799 (période d'essai de notre test qui est incluse dans nos relevés actuels) : $N = 1446$ pour 1693 actes relevés. La probabilité maximale de trouver une correspondance est donc $Pr_{max} = 0.716$, en se restreignant au cas où l'âge au décès est donné par le prêtre.

[5]Programme d'exploitation des tables et bases de données.
[6]Ce dernier problème est levé en intégrant le calcul de e dans un voisinage $\vartheta\left(age_{DC}\left(y\right)\right)$
[7]Ces dates figurent dans les actes de décès

(2) Dans une deuxième phase, pour tout individu $y = \{nom,\ \cdot,\ sexe,\ date_{dc},\ age_{dc}\}$ de nom, sexe, âge au décès donné, prenom, on cherche à vérifier les relations :

$$\begin{aligned} nom\,(y) = nom\,(x) \quad &\Leftrightarrow \quad C_1 \\ sexe\,(y) = sexe\,(x) \quad &\Leftrightarrow \quad C_2 \\ e < a\ vrai \quad &\Leftrightarrow \quad C_3 \\ prenom\,(y) = prenom\,(x) \quad &\Leftrightarrow \quad C_4 \end{aligned}$$

on utilise l'âge au moment du décès : e et introduit le critère d'âge :

$$e = \frac{(date_{dc}(y) - date_{naissance}\,(y)) - age_{dc}}{age_{dc}} < a \tag{4.3}$$

(3) On fait varier a de 0.03 à 0.5

(4) Enfin, pour chaque valeur de a, on totalise le nombre d'éléments (n) tels que les $n - uplets$ de x et y coïncident :

$$u = f\,(x, y)\,,\ f = \{0, 1\}\ \ ou\ u = g\,(x, y)\,,\ g = \{0, 1\}$$

Si $f = 1$ alors $C_1 \Leftarrow nom\,(y) = nom\,(x)$ et $C_2 \Leftarrow sexe\,(y) = sexe\,(x)$, de plus $e < a \Rightarrow C_3$ (eq : 4.3) et $prenom\,(y) = prenom\,(x) \Rightarrow C_4$; la sélection maximale vérifie C_1 , C_2, C_3 et C_4 et renvoit de $\mathcal{B}$ à S_k, la sélection minimale vérifie C_1 , C_2, C_3 renvoit de $\mathcal{B}$ à S_k.

si $f = 0$ aucune des conditions n'est vérifiée.

Il en va de même avec la fonction g qui renvoit une partie de $\mathcal{B}$ dans S_u, mais la condition C_3 n'existe pas !

Remark 4.3. Tous les éléments de la table des décès $\mathcal{S}$ ne sont pas classés.

Ceux qui sont classés appartiennent à S_k(âge au décès connu) ou à S_u (âge au décès non donné).

$S_k$et $S_u$sont inclus dans S

$\overline{S} = C_S\,(S_k \cup S_u)$ le complément de $S_k \cup S_u$ dans $\mathcal{S}$ représenterait donc l'ensemble des individus décédés dans la paroisse sur la période donnée, mais nés en dehors de la paroisse.

L'outil ainsi développé pourrait nous permettre d'étudier la démographie d'un village sous l'ancien régime ?

La figure ci-dessous (fig : 4.1) donne un graphe de correspondances $y \rightarrow x$

> ***Nota bene :*** *Le nombre d'éléments $x_i\,, i = 0 \cdots m \in \mathcal{B}$ pouvant correspondre à un élément $y \in \mathcal{S}$ varie de zéro (inconnu à la naissance) à plusieurs unités (attribution impossible par le test).*

4.4. **Résultats :**

(1) Le nombre total n de personnes ainsi classées de date de décès connue, ayant au moins un antécédent de même nom dans la base de naissances (y possèdant une image i.e. $f = 1$) varie de $n = 506$ pour $a = 3\%$ à $n = 770$ pour $a = 50\%$

(2) Le nombre de personnes n classées donnant une correspondance unique (prénoms égaux à une altération d'orthographe près C_4vrai, $y \in S_k$) est de 255 pour $a = 3\%$,

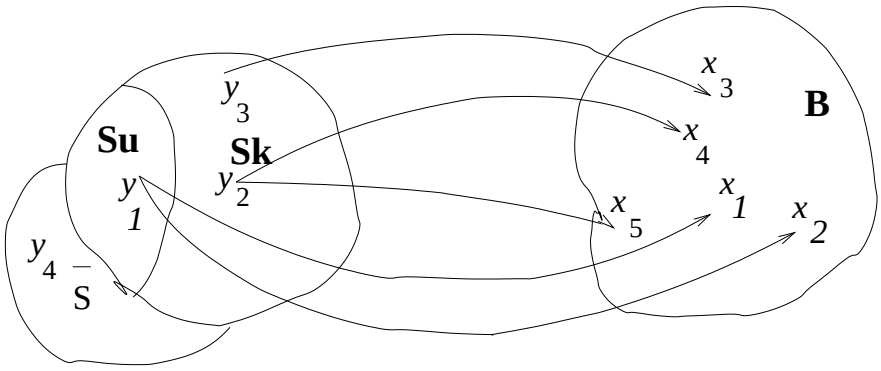

FIG. 4.1. Les parties de l'ensemble des individus décédés à Mussy et leurs relations avec les naissances

(3) Enfin, le nombre total de correspondances $y \in S_k \cup S_u$pour $a = 50\%$ est de 1050 (critère à 50%). Ce nombre avoisine le nombre d'individus dont l'âge est connu au moment du décès (1446). Ceci justifie, selon nous, l'arrêt de l'augmentation de a (on atteint 1079 pour $a = 75\%$),

5. ANALYSE DES RÉSULTATS

5.1. **Sélection minimale.** On définit la sélection minimale comme il a été dit ci-dessus (sans le prénom) :

$$pour\ y = \{nom,\ \cdot,\ sexe,\ date_{dc},\ age_{dc}\}\ \exists x = \{nom,\ \cdot,\ sexe,\ date_{naiss}\}\ tq\ u \neq \varnothing$$

Remark. Un individu décédé à Mussy entre 1750 et 1799 a de fortes chances d'être né à Mussy.

Un individu né à Mussy avant 1750 peut être vivant après 1800, il y a probablement plus de 50% de chances qu'il décède à Mussy.

Il est intéressant de noter que le fait d'augmenter le taux d'erreur dans la détermination de l'âge par le prêtre de 5 à 50% entraine presque le doublement du nombre de correspondances décès-naissance résolues par la fonction inverse (on passe de 618 à 1036). Un individu qui serait mort à l'âge estimé de 40 ans (espérance de vie milieu du XVIIIè siècle) verrait son âge réel situé entre 39 et 41ans dans la détermination à 5% et entre 30 et 50 ans dans le second cas.

Le faible recouvrement des décès par les naissances (nombre d'éléments en correspondance avec $a = 3\%$) serait peut-être dû aux erreurs d'orthographe dans l'écriture des prénoms, au changement du prénom donné au décès, à une approximation trop grande de l'âge du défunt relevé par le prêtre au décès ?

La correspondance obtenue est meilleure... mais ***reste assez faible***. Nous pouvons regrouper ultérieurement les individus par nom et prénom (altéré ou modifié).

Example 5.1. *Illustrons les cas de prénoms altéré et modifié. Prénom altéré non trouvé lors de la recherche précédente : Desmurger* ***Benoist*** *né le 8 janvier 1714 devient* ***Benoit*** *à son décès le 20 avril 1778, âgé de 66 ans ($a < 5\%$). Desmurger* ***Pierre*** y_1 *décédé le 27 mars 1776 âgé de 50 ans correspond à* ***Guillaume (***x_1***)*** *né le 14 novembre*

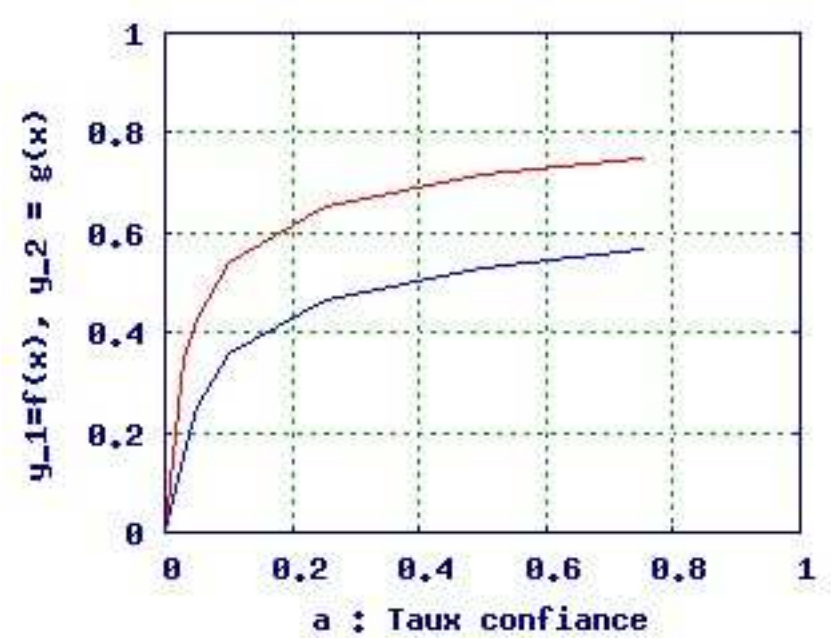

FIG. 5.1. Evolution de l'identification en fonction du taux de confiance sur l'âge au décès (a)

1725, un autre individu correspondant approximativement à l'âge pourrait être ***Jean (***x_2***),*** *né le 10 février 1724... dans ce dernier cas pour* $y_1 \leftarrow \{x_1, x_2\}$

Nous trouvons environ 830 **individus en correspondance d'âge et de sexe, avec un taux de confiance à** $a < 5\%$**, au moment du décès. Le nombre d'individus correctement isolés des tables avec l'orthographe des prénoms altérée passe de** 360 **à** 618 **dans** S_k**, le nombre de personnes identifiées dans** S_k **avec la concordance du nom, l'âge et le sexe est de** 830.

$$card\,(E_{dc}) = 830,\ Pr_{0.95} = \frac{830}{1693} \simeq 49\%$$

La figure (fig : 5.1) donne l'évolution du taux d'identification des individus avec une sélection maximale (resp : minimale). On voit que pour atteindre une bonne précision dans nos estimations, il faut fixer $a \simeq 0.3$ Au delà de cette valeur le taux n'augmente que faiblement et la marge d'erreur sur l'âge augmente énormément.

Le tableau suivant (tab : 3 on the next page) donne une liste des cas résolus et restant à résoudre pour les familles les mieux représentées à Mussy sous Dun.

Les résultats présentés dans le tableau ci-dessus montrent bien que 25 familles représentent 66% de la population ($\frac{465}{700}$).

5.2. **Analyse dans un voisinage à 0.3.** Nous avons donc repris la recherche des noms dans la liste de 1693 décès survenus dans la période test : $1700 - 1749$, en fixant la valeur du test $|e| < 0.3$ afin d'identifier environ 70% de la population.

(5.1) $$e \in \vartheta\,(age_{DC}\,(y))$$

TAB. 3. Liste des principales familles dont les individus sont identifiables (de 1738 à 1812)

Nom	cas résolus $y \leftarrow x$	restant à préciser $y \leftarrow \{x_1, \cdots, x_k\}$	Qualité de l'estimation $\%$
Auberger	26	16	62
Augagneur	6	5	55
Bajard	27	26	51
Ballandras	8	6	57
Bidaud	44	13	77
Chemier	28	8	78
Chizelle	42	17	71
Corneloup	15	7	68
Desmurger	21	5	80
Duris	15	6	71
Jolivet	1	6	14
Labrosse	23	19	55
Lachize	11	18	38
Lacombe	27	5	84
Laroche	18	13	58
Lathuillière	10	7	59
Livet	17	14	55
Martin	38	10	79
Millet	30	6	83
Troncy	11	8	58
Trouillet	27	7	79
Verchère	20	8	71
Total :	465	230	

TAB. 4. Résultats d'identification avec $|e| < 0.3$

Sous-espace	S_k	S_u	$\overline{\mathcal{S}}$	$\mathcal{S}$	non classés
Sélection nom, prénom	884	140		1693	669
Sélection nom seul	1069	424	30	1693	170
total classés (%) :	63%	25%	1.7%	100%	10.3%

Les valeurs obtenues sont reportées dans le tableau suivant (tab : 4).

Les résultats sont assez satisfaisants puisqu'environ 50%de la population décédée dans cette tranche est identifiée , (63%) si l'on suppose que les prénoms peuvent changer entre l'acte de naissance et celui de décès !

6. CONCLUSIONS

Le recouvrement des tables de naissance - décès reste faible $Pr_{0.95} \simeq 50\%$, on s'attendrait à trouver beaucoup plus.

Faut-il conclure que les prénoms des individus pouvaient varier entre la naissance et le décès ?

Il est a peu près certain que de nombreuses erreurs ou des défauts d'attribution peuvent être commis par les généalogistes amateurs, lorsqu'ils cherchent à « suivre » leurs ancêtres dans une lecture hâtive des archives !

Nous poursuivrons ce travail sur les périodes $1750-1799$, $1800-1849$, $1850-1899$.

Nous tenons à disposition des personnes intéressées le tableau des derniers résultats que nous jugeons les plus aptes à être exploités pour relever correctement nos ancêtres à Mussy sous Dun[8].

[8]Bien qu'il subsiste de nombreuses erreurs dans les relevés des traducteurs, nul n'est parfait, le travail n'est pas facile.