Traitement des données des trajectoires de vie de l'enquête 3B réalisée par l'INED en 1981

Cahier des charges

BENKAOUR Salwa

FLEURY Pierre

JORDAN Célia

Table des matières

Table des illustrations

1 Introduction

Le projet est réalisé par trois étudiants du master MIASHS (Mathématique et Informatique Appliqués aux Sciences Humaines et Sociales) parcours IC (Informatique et Cognition), il a lieu dans un cadre pédagogique.

Ce document a pour objectif de mettre au clair les attentes de notre projet en termes de réalisation et les moyens que nous mettons en œuvre pour y parvenir.

Le projet se base sur une enquête qui a été réalisée par l'INED en 1981. Il permet d'étudier les données qui ont été récoltées par cette enquête et d'analyser le comportement des individus par rapport à leur parcours de vie. On veut mieux comprendre les choix et les motivations des individus au cours de leur vie.

Il est important de comprendre les notions qui vont être utilisées. En effet, la notion de trajectoire de vie est complexe. C'est un objet d'études qui sert de support à l'analyse des motivations qui déterminent les choix d'un individu dans son parcours biographique. C'est-à-dire en mettant en perspective dans le temps les informations disponibles sur les individus pour mieux comprendre et expliquer ce qui déterminent leurs choix et, par conséquent, leurs parcours.

On cherche un modèle pour représenter des trajectoires de vies données.

Des modèles existent déjà mais on vise plus de généricité ici, avec la prise en compte d'un modèle multidimensionnel et de facteurs explicatifs autres que ceux déjà rapportés dans des travaux. (Thériault, par exemple, cherche à analyser la dimension résidentielle des trajectoires de vie.)

Pour réaliser le modèle, un fichier csv nous est fourni, il contient des données, récoltées par l'INED, qui sont à analyser. C'est un document qui contient les réponses à une enquête réalisée en 1981. Cette enquête est effectuée sur une population âgée de 45 à 69 ans à l'époque afin d'étudier 3 grandes dimensions ; la dimension familiale, résidentielle et professionnelle. Le document csv contient les types de données qui suivent :

V002	Année	DepCom	Nom_ commune	Pers	Type.Lieu	Rang	TU
Numéro de l'enquêté	**Année de l'évènement /l'action**	**Numéro de commune**	**Nom de la commune**	**Cjt :** Conjoint	**Att :** Attache (les séjours fréquents)	**Nb de fois que cela s'est produit.**	**Type d'unité, codification par rapport à la commune**
				Ego : Enquêté	**Etu :** Etude		
				Enf : Enfant	**Mar :** Mariage		
				Par : Parent	**Nais :** Naissance		
					Pens : Pensionnaire		
					Proj : Projet		
					Res : Résidence		
					Sejour : Vacances		
					Sep : Sépulture (décès)		
					Work : Travail		

FIGURE 1 : DESCRIPTION DES LIGNES ET DES COLONNES

Ces données ont été fournies par l'INED suite à la thèse de M. David Noel qui a fait une étude sur le cas de la trajectoire résidentielle.

Le but de ce projet est d'analyser les données qui nous ont été fournies et si le temps le permet d'accéder à leur visualisation.

2 Guide de lecture

2.1 Maîtrise d'œuvre

2.1.1 Responsable

Jérôme Gensel dirige nos travaux.

2.1.2 Personnel administratif

Notre contact pour les détails administratifs du stage du côté du Laboratoire Informatique de Grenoble (LIG) est Laurence Schimicci.

2.1.3 Personnel technique

L'équipe autour du projet est composée de Marlène Villanova-Oliver, Camille Bernard et Jérôme Gensel. Un doctorant rejoint cette équipe en novembre.

2.2 Maîtrise d'ouvrage

2.2.1 Responsable

Damien Pellier est le responsable administratif du groupe pour le projet. Nous le voyons toutes les semaines pour rendre compte de l'avancement du projet.

2.2.2 Personnel administratif

Notre contact pour les détails administratifs du stage du master MIASHS parcours IC, de l'UFR SHS de l'Université Grenoble Alpes (UGA) est Fatima Belounis.

2.2.3 Personnel technique

Le groupe est composé de Salwa Benkaour, Pierre Fleury et Célia Jordan. Nous sommes trois étudiants du Master Mathématiques et Informatique Appliqués aux Sciences Humaines et Sociales mention Informatique & Cognition (MIASHS - IC).

3 Concept de base

Trajectoires de vie : Les trajectoires de vie sont au centre du sujet. On les désigne comme des objets qui permettent d'analyser la vie d'un individu. Grâce à ces dernières, on cherche à analyser les changements qui surviennent dans la vie d'un individu à un moment donné de sa vie. Le temps est donc une notion centrale dans ce concept. Hélardot (2006) définit les trajectoires de vie comme « un entrecroisement de multiples lignes biographiques plus ou moins autonomes ou dépendantes les unes des autres ». Les lignes de vie sont alors des domaines de l'existence. Dans ce sujet les domaines auxquels nous nous intéressons sont les domaines familiaux, professionnels, résidentiels et de voyage.

Une trajectoire peut être considérée comme composée de plusieurs sous-trajectoires. Toutes ont en commun de posséder des *épisodes* et des *événements*. Le premier représente une période stable et le second un changement entre deux de ces périodes. A noter que chaque épisode commence et se termine par un événement.

Ce modèle multipoint de vue a été développé par David Noël en 2019 et se nomme LTOP (voir figure 1), *Life Trajectory Ontlogy Pattern*, et sert à modéliser des trajectoires de vies composées de sous-trajectoires.

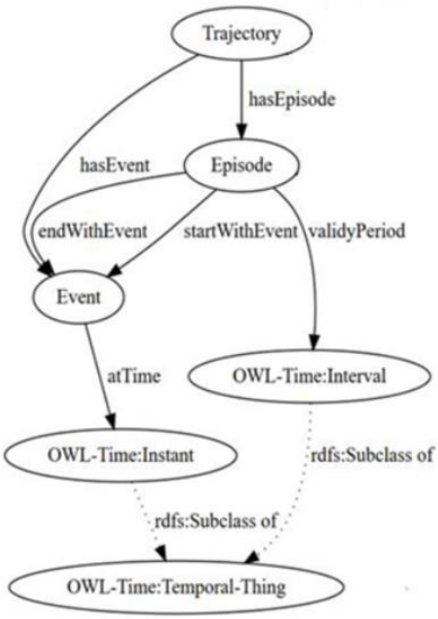


FIGURE 2 : PATRON DE CONCEPTION LTOP EXTRAIT DU SUJET

Enfin, le modèle LTOP prend en compte différents niveaux de granularité. Ces derniers déterminent la précision de l'information à disposition. L'explication peut donc varier selon ces niveaux.

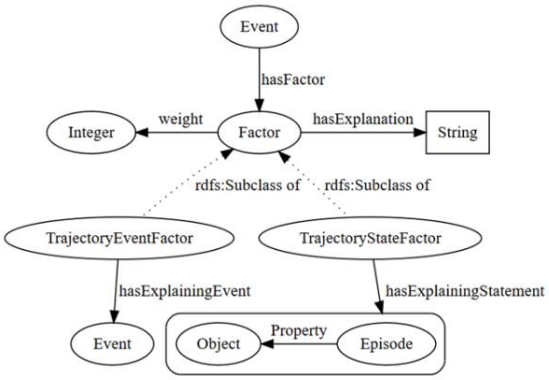


FIGURE 3 : MODELE DE FACTEUR EXPLICATIF EXTRAIT DU SUJET

Facteur explicatif : La notion de facteur explicatif cherche simplement à donner des explications cohérentes aux événements ou aux épisodes. Ces explications sont internes si l'explication appartient à la même trajectoire et externe sinon. Toutes les trajectoires ont en commun le temps, ce qui permet de croiser les données et trouver des facteurs explicatifs. Il est indiqué que l'espace peut aussi être commun à plusieurs trajectoires, mais pas à toutes. Les facteurs explicatifs sont donnés par l'individu.

Multi-dimensionnalité : On utilise la multi-représentation pour connaître différents points de vue sur une entité donnée. De plus, d'après David Noël, "une représentation est la partie d'un modèle qui correspond à un point de vue particulier". Il existe plusieurs approches pour la modélisation de multiples représentations dont l'UML (cf notre modèle).

Multi-granularité induite : La granularité de l'information définit le détail que nous apportons à cette information. L'enquête 3B a été réalisée de telle manière qu'on demande aux enquêtés des souvenirs, c'est-à-dire des approximations de la réalité. De ce fait, il est important de définir des niveaux de granularité pour nous permettre d'avoir une "marge". De la même manière, le fait que nous

envisageons de répliquer l'analyse sur d'autres données nous oblige à être plus conciliant quant à la précision de nos informations. On peut avoir une meilleure exploitation des données en jouant sur la granularité. Pour ce qui est de l'information spatiale, il est important de prendre en compte quelques niveaux de granularité pour la simple raison que les temps changent mais aussi les espaces. L'enquête ayant environ 40 ans, beaucoup de changements ont eu lieu sur le territoire, dans les villes et villages, ne serait-ce que par leur agrandissement, dûs à une population émergente. D'après Vangenot et al. (2002), les niveaux de granularité sont des perceptions, ce qui nous permet de représenter cette granularité par la multi-représentation, définie ci-dessus. En ce sens, les niveaux de granularité sont représentés par les points de vue du modèle, ici la famille, le travail, la résidence.

4 Contexte

Notre projet se base sur une étude faite par sondages et prise de notes manuscrites, qui a été réalisée en 1981 par l'institut national d'études démographiques (INED). C'est un questionnaire nommé « questionnaire 3B » pour 3 biographies. En effet, il y a la biographie familiale, professionnelle et migratoire. Ce questionnaire à été réalisé sur une population âgée de 45 à 69 ans. Nous n'avons pas toutes les données qui ont été demandées à l'enquêté, toutefois, le fichier qui nous est fourni contient 10 colonnes et 493 856 lignes. C'est un fichier réalisé au format csv ageven, ce qui correspond respectivement à "age" pour l'âge est "even" pour l'événement, on attribue ainsi une date à chaque événement. On a un seul élément "even" par ligne "age" dans notre fichier.

David Noël a produit une thèse concernant les trajectoires de vie et plus particulièrement le cas de l'étude de la trajectoire résidentielle. Pour cela il a utilisé un patron d'une ontologie de trajectoire de vie nommée LTOP (Life Trajectory Ontology Pattern), ce qui lui permet de représenter les différentes dimensions des lignes de vie.

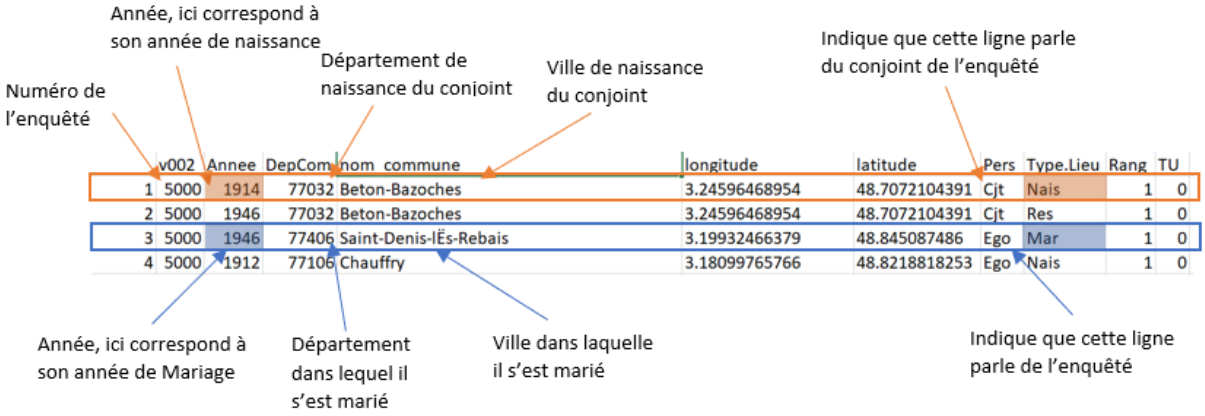
Pour notre projet, nous n'utilisons pas ce qui est relatif aux données RDF, initialement prévues. Nous devons substituer cette approche Web Sémantique par une approche SGBD relationnelle en créant une base de données relationnelle dont le schéma devra s'inspirer du modèle de trajectoire de vie de David Noël. Une fois cette base de données remplie par les données du fichier 3B.csv ces données extraites pourront être exploitées via des requêtes SQL ou/et via un ou des programmes Java permettant par exemple, leur visualisation. La base de données sera remplie grâce à un algorithme que nous devons écrire et qui doit être en accord avec la thèse générique de David Noël, pour cela nous utiliserons le Modele_3B.

5 Description de la demande

5.1 Les objectifs

Nous devons créer un modèle UML des lignes de vies, et à partir de celui-ci, produire une base de données relationnelle.
L'obtention de cette base de données peut se faire de deux façons différentes.
Soit en utilisant un parseur et ainsi on crée la base de données à partir du fichier csv.
Soit en utilisant une autre base de données qui regroupe toutes les informations du fichier csv que l'on interroge afin d'obtenir une base de données en accord avec notre modèle.

On décrit ci-dessous la nature des données qui nous ont été fournies grâce à l'enquête 3B.



	v002	Annee	DepCom	nom_commune	longitude	latitude	Pers	Type.Lieu	Rang	TU
1	5000	1914	77032	Beton-Bazoches	3.24596468954	48.7072104391	Cjt	Nais	1	0
2	5000	1946	77032	Beton-Bazoches	3.24596468954	48.7072104391	Cjt	Res	1	0
3	5000	1946	77406	Saint-Denis-lÈs-Rebais	3.19932466379	48.845087486	Ego	Mar	1	0
4	5000	1912	77106	Chauffry	3.18099765766	48.8218818253	Ego	Nais	1	0

FIGURE 4 : EXPLICATION DES LIGNES

Dans la figure ci-dessus, la première ligne correspond à un événement concernant l'individu 5000. Cet événement se déroule en 1914 et concerne le conjoint de la personne interrogée. Plus précisément, cette date correspond à la naissance de son conjoint, qui a lieu dans la ville de Beton-Bazoches dans le 77.
La deuxième ligne correspond au lieu de résidence du conjoint de l'enquêté en 1946, soit la ville de Beton-Bazoches dans le 77.
La troisième ligne correspond au mariage de l'enquêté cette fois-ci, toujours le numéro 5000, dans la ville de Saint-Denis-lès-Rebais en 1946 dans le 77 en 1946.
La dernière ligne correspond à la naissance de l'enquêté, qui a lieu dans la ville de Chauffry dans le 77 en 1912.

La longitude et la latitude sont les données de ville correspondantes. Elles permettent de situer géographiquement l'événement et donc nous pouvons intercaler cela avec une analyse spatiale.
Le rang nous donne le nombre de fois que l'événement s'est produit et TU nous donne l'unité territoriale.

Nous pouvons constater qu'il est possible d'inférer un certain nombre d'événements qui ne sont pas relatés directement dans ces données. Chaque changement est un événement et chaque période est un épisode. Ainsi, le premier épisode d'une personne commence par sa naissance par exemple.
De plus, on peut introduire par l'exemple les avantages des niveaux de granularité. Ici, on possède l'information relative au département, ce qui peut nous aider dans l'analyse des déplacements sur le territoire en fonction de la situation maritale par exemple. Peut-être qu'avoir une famille entraîne une restriction des mouvements de manière globale. Cette hypothèse est difficile à analyser car, ne l'oublions pas, il faut garder à l'esprit que l'enquête est réalisée en 1981 et qu'on discute de ce qu'il s'est passé bien avant cette date. Pour l'exemple, les moyens de locomotions sont bien moins développés à l'époque qu'aujourd'hui.

Ainsi, à l'aide du modèle UML (cf. ressources) on retient le dictionnaire suivant :

Propriété	Type	Description
id	entier	id de la personne
annee	entier	année de l'événement
dep	texte	département de l'événement
commune	texte	commune d'habitat au moment de l'événement
longitude	double	longitude de la commune
latitude	double	latitude de la commune
personne	texte	personne concernée par l'événement

type_lieu	texte	type de lieu de l'événement
rang	texte	nombre de fois que l'événement en question s'est produit
evenement	texte	description de l'événement
episode	texte	description de l'épisode
localisation	texte	description de la localisation de l'événement
rang	texte	nombre de fois que l'événement en question s'est produit
mere	texte	Mère de la personne interrogée
pere	texte	Père de la personne interrogée
enfant	texte	Enfant de la personne interrogée
conjoint	texte	Conjoint de la personne interrogée
f_trajectory	abstrait	trajectoire familiale
f_episode	texte	description de l'épisode
f_event	texte	description de la localisation de l'événement
p_trajectory	abstrait	trajectoire professionnelle
p_episode	texte	description de l'épisode
p_event	texte	description de la localisation de l'événement
l_trajectory	abstrait	trajectoire de loisirs
l_episode	texte	description de l'épisode
l_event	texte	description de la localisation de l'événement
r_trajectory	abstrait	trajectoire résidentielle
r_episode	texte	description de l'épisode
r_event	texte	description de la localisation de l'événement
f_event	texte	description de la localisation de l'événement

On veut pouvoir répondre à des questions comme :

- Combien de fois l'individu X a changé d'emploi ?

- Combien d'enfants à l'individu Y ? etc…

La construction de requêtes nous permettra de répondre à ces questions.

5.2 Produit du projet

Le produit est une base de données relationnelle qui devra être exploitée par SQL ou par des programmes Java. Des requêtes pourront ensuite lui être posées afin de pouvoir étudier certains comportements. Idéalement, le produit doit pouvoir nous donner une interface qui permettra de mieux visualiser les informations qu'il contient. On doit pouvoir exploiter les données obtenues afin de les étudier.

5.3 Les fonctions du projet

Le produit doit pouvoir fonctionner avec une autre enquête du nom de EDER qui est elle aussi disponible au format csv ageven pour réfléchir à une généricité de l'approche. Il doit pouvoir ajouter de nouveaux éléments mais aussi en supprimer. On doit aussi pouvoir modifier ce que l'on a déjà, si par exemple on a de nouvelles informations.

5.4 Critères d'acceptabilité et de réception

Le produit doit correspondre au modèle UML choisi, c'est-à-dire au modèle 3B tout en se basant sur le modèle LTOP. Si celui-ci ne correspond pas à l'un des deux, il n'est pas acceptable.

6 Contraintes

6.1 Contraintes de coûts

A priori il n'y a aucun coût pour nous.

6.2 Contraintes de délais

Le stage à plein temps a lieu de mi-mai 2022 jusqu'au 30 juin 2022. C'est à la fin de cette période que le produit doit être fourni. On entend par produit la création de la base de données et les requêtes qui permettent d'exploiter les tables.

6.3 Contraintes matérielles

Pour faire fonctionner le produit, l'utilisateur doit bien sûr avoir un ordinateur fonctionnel.

6.4 Autres contraintes

Contraintes juridiques sur les données ? Les données 3B sont fournies par l'INED à l'équipe et nous sont prêtés lors du stage. Nous n'avons pas le droit de les utiliser en dehors du cadre du stage.

Contraintes logicielles ? Nous utiliserons Postgresql comme SGBD car il contient l'extension PostGIS qui permet d'avoir une représentation spatiale de ses données.

7 Déroulement du projet

7.1 Planification

Le projet se découpe en deux grandes parties. La première a lieu d'octobre 2021 à janvier 2022 et consiste en la planification de ce qui sera fait durant la seconde période. Cette dernière consiste au développement du produit.

La première phase débouche sur une soutenance intermédiaire qui a lieu le 17 ou 18 janvier 2022. La seconde phase se termine par une soutenance le 23 juin 2022.

- algo 3B

- plan de sondage

- grille recueil lieux 3B

- données 3B

- 3 questionnaires

- dépliant ined insee

- instruction générale sur le déroulement de l'enquête

- quelques remarques sur la présentation du questionnaire

- présentation de l'enquête 3B

- bibliographie de l’enquête

8 Références

Noël, D. (2019). *Une approche basée sur le web sémantique pour l’étude de trajectoires de vie*.

9 Index