République Algérienne démocratique et Populaire
Ministère de l'enseignement supérieur et de la recherche scientifique
UNIVERSITE Dr.TAHAR MOULAY SAIDA
FACULTE : TECHNOLOGIE
DEPARTEMENT : INFORMATIQUE

MEMOIRE DE FIN D'ETUDES EN VUE DE L'OBTENTION DU DIPLOME DE MASTER EN INFORMATIQUE

OPTION : Réseaux Informatique et Système Répartie

Thème

UTILISATION DU COUPLAGE D'ENREGISTREMENTS POUR LA QUALITE DE DONNEES

Présenté par :
BELHADJ Imane .
AMARA Zohra.

Encadré par :
Mr. BENYAHYA Miloud
Mr. BENYAHYA Kadda

Promotion Septembre 2020

RÉSUMÉ

La qualité des données dans les bases, les entrepôts de données ou plus généralement dans les systèmes d'informations est un enjeu majeur.

De plus en plus d'applications utilisent les données en ligne, or ces données souffrent aujourd'hui d'un manque de fiabilité : erreurs, données isolées, doublons, incohérences, valeurs manquantes, incomplètes, incertaines, obsolètes, ou peu fiables. Les problèmes liés à la qualité des données sont coûteux et universels. Le couplage d'enregistrements est la tâche de trouver des enregistrements dans un ensemble de données qui se réfèrent à la même entité dans les différentes sources de données. Le but de notre travail est d'utilise le couplage probabiliste d'enregistrements pour résoudre ce problème tout en se basant sur la combinaison de plusieurs mesures de similarité entre attributs.

Mots clés : Qualité des données, Couplage d'enregistrement, mesures de similarité.

Abstract

The quality of data in databases, data warehouses or more generally in information systems is a major issue. More and more applications are using online data, and this data today suffers from a lack of reliability: errors, isolated data, duplicates, inconsistencies, missing values, incomplete, uncertain, obsolete, or unreliable. Data quality issues are costly and universal. Record linkage is the task of finding records in a data set that refer to the same entity in different data sources. The aim of our work is to use probabilistic record linkage to solve this problem while relying on the combination of several measures of similarity between attributes

Keywords: Data quality, Record linkage, similarity measures.

Remerciement

Tout travail réussi dans la vie nécessite en premier lieu la Bénédiction de Dieu, et ensuite l'aide et le support de plusieurs personnes. Nous tenons donc à remercier et à adresser notre reconnaissance à toute personne qui nous a aidés de loin ou de près afin de réaliser ce travail.

Nous exprimons ici notre profonde reconnaissance à l'égard de notre promoteur Mr. Miloud BENYAHYA . Il a su nous orienter dans la recherche de notre thème. Les conseils et les encouragements qu'il nous a jamais cessé de prodiguer sont inestimables. Sa patience et sa compréhension nous ont permis d'avancer et de terminer notre mémoire.

Dédicace

Tous les mots ne sauraient exprimer la gratitude, l'amour, le respect, la reconnaissance, c'est tous simplement que : Je dédie cette mémoire de master à :

A Ma tendre Mère Safia : Tu représente pour moi la source de tendresse et l'exemple de dévouement qui n'a pas cessé de m'encourager. Tu as fait plus qu'une mère puisse faire pour que ses enfants suivent le bon chemin dans leur vie et leurs études.

A Mon très cher Père Mansour : Aucune dédicace ne saurait exprimer l'amour, l'estime, le dévouement et le respect que j'ai toujours pour vous. Rien au monde ne vaut les efforts fournis jour et nuit pour mon éducation et mon bien être. Ce travail et le fruit de tes sacrifices que tu as consentis pour mon éducation et ma formation le long de ces années.

A mon très cher mari zakaria bouzad : Tes sacrifices, ton soutien moral et matériel m'ont permis de réussir mes études. Ce travail soit témoignage de ma reconnaissance et de mon amour sincère et fidèle.

A mes frères walid, nasro .

A mes sœurs : Rachida, Houria et ses enfants Mohamed et Fatima , Hidayat et ses enfants Mansour et Hadjer..

A mes chère belle sœurs : Hanen, hayat, ikram, djamila , amira , nadia, marwa bouzad .

Mes vifs remerciements vont également à ma binôme :Amara Zohra A mes très chère amis : Ahlem, bekhta; madjda imane

A mes oncles et leurs enfants, À mes tantes et leurs enfants et a toute la famille BELHADJ et la famille BOUZAD

. Cette humble dédicace ne saurait exprimer mon grand respect et ma profonde estime.

A tous les membres de ma promotion.

A tous mes enseignants depuis mes premières années d'études.

Imane

Dédicace :

A mes chers parents, pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de mes études,
A mes chères sœurs makhlofia, imane pour leurs encouragements permanents, et leur soutien moral,
A mon cher frère, abd lmalek pour leur appui et leur encouragement,
Mes vifs remerciements vont également à ma binôme :Belhadj Imane et à tous mes amies intimes: *bekhta , sara ,maria*
A toute ma famille pour leur soutien tout au long de mon parcours universitaire,
Que ce travail soit l'accomplissement de vos vœux tant allégués, et le fuit de votre soutien infaillible,
Merci d'être toujours là pour moi

Zohra

TABLE DES MATIÈRES

TABLE DES FIGURES

INTRODUCTION GÉNÉRALE

Pour surmonter les problèmes de la qualité des données résultant de l'intégration de données provenant de diverses sources réparties, autonome et hétérogène tels que les erreurs typographiques et les variations, des pistes prometteuses sont ouverts et des différentes techniques sont utilisées.

Le Couplage d'enregistrements également connu sous le nom de résolution d'entité, de correspondance de données et de détection doublon, permet d'identification et de mise en correspondance d'enregistrements de mêmes entités du monde réel partir de différentes sources. [6]

Nous allons intéresser au problème d'identification des enregistrements qui faire référence aux mêmes entités du monde réel, ces enregistrements sont liés sur la base de la similitude entre attributs communs, chaque paire étant classée comme lien ou non-lien en fonction de leur similitude.

Les techniques de comparaison de chaînes approximatives sont utilisées pour comparer les paires d'enregistrements, ce qui conduit à un vecteur de similitudes pour chaque paire. Ils sont utilisés pour classer les paires d'enregistrements en liens et non-liens. [6]

Le couplage probabiliste d'enregistrements pour résoudre ce problème tout en se basant sur la combinaison de plusieurs mesures de similarité entre attributs.

Le mémoire est structuré en 4 chapitres :

Dans le premier chapitre : Nous allons présenter la qualité des données, nous donnerons un aperçu sur la qualité et nous parlerons sur les critères qui définissent la qualité des données, Les objectifs et les problèmes de la non-qualité des données.

Dans le deuxième chapitre : Nous Parlerons de couplage d'enregistrement et nous présentons les différents travaux et approches utilisés dans ce contexte puis nous détaillons les algorithmes proposés et nous comparons les résultats obtenus.

Dans le troisième chapitre : Nous présentons l'approche de résolution d'entités via le couplage d'enregistrements et nous avons proposés les techniques de comparaison d'attributs et voir les différentes manières de combinaisons les mesures de similarité.

Dans le quatrième chapitre : Nous décrivons l'environnement de l'implémentation ensuite nous présentons notre application où nous évaluons Algorithme de couplage d'enregistrements sur des bases de données réelles et nous terminons par les tests effectues et les résultats obtenus.

CHAPITRE 1

LA QUALITÉ DES DONNÉES

1.1 Introduction

De nos jours, les données représentent une richesse pour les entreprises et les administrations et contribuent à leur développement. La qualité de ces données représente un enjeu important.

Le coût de la non-qualité peut en effet s'avérer très élevé : prendre une décision à partir de mauvaises informations peut nuire à l'organisation, à ses clients ou ses partenaires. La gouvernance des données est un sujet qui prend de l'importance dans les entreprises et les administrations. Elle permet l'amélioration des interactions entre les différents collaborateurs d'une ou plusieurs organisations concernées [1]

Dans ce chapitre, nous commencerons par définir qualité des données et leur concept, par la suite on abordera la problématique de la qualité de données. Ensuite on citera les dimensions et Les approches génériques, Les approches spécifiques de la qualité des données. Et finalement on parlera sur les objectifs de qualité et Coût du non qualité.

1.2 La qualité de donnée

La qualité est une préoccupation que l'on trouve dans beaucoup de domaine De ce fait la première difficulté réside dans l'absence de consensus sur la notion de qualité Comme la communauté aujourd'hui préconise également l'application dés le début des normes et standards internationaux, nous intéressons ici aux définitions données par l'organisation internationale de standardisation (ISO : International StandarOrganization) et par Organisation de Coopération et de Développement Economiques (OECD : Organisation for Economic Cooperation and Development).[2]

ISO : définit la qualité comme L'ensemble des propriétés et caractéristiques d'un produit ou d'un service qui lui confère l'aptitude a satisfaire des besoins exprimes ou impliquâtes Pratiquement, la qualité d'un produit signifie qu'il est adapte au besoin qu'il est censé satisfaire. La notion de qualité s'applique aussi bien a des produits qu'a des services [1]

OECD : La qualité est vue comme un concept à facettes multiples. Les caractéristiques de qualité dépendent des perspectives, des besoins et des priorités d'utilisateur, qui changent à travers des groupes d'utilisateurs. Ainsi cette définition est complémentaire à la définition ISO en y ajoutant le contexte d'utilisation et le domaine de l'application c.à.d. que les besoins sont définis par l'utilisateur dans le cadre d'une application donnée. [1]

1.2.1 Définition La qualité des données

La qualité des données est un terme générique décrivant à la fois les caractéristiques de données : complètes, fiables, pertinentes, à jour et cohérentes, mais aussi l'ensemble du processus qui permet de garantir ses caractéristiques. Le but est d'obtenir des données sans doublons, sans fautes d'orthographes, sans omission, sans variation superflue et conforme à la structure définie

1.2.2 Types de données et systèmes d'information :

De façon générale, nous remarquons que la problématique de la qualité de données concerne principalement les dimensions qualité (i.e. précision, exactitude, cohérence, etc.), les modèles, les techniques, les

outils, ainsi que les méthodologies adaptées aux nouveaux types de données et systèmes d'information [4] :

· ***Les dimensions :*** sont normalement appliquées dans les modèles qualité, ainsi que dans les techniques, les outils et les structures.

· ***Les modèles :*** généralement utilises dans les bases de données, ont été enrichis afin de représenter les dimensions et autres aspects lies a la qualité des données.

· ***Les techniques :*** sont un ensemble d'algorithmes, d'heuristiques, et de processus pour répondre a un problème spécifique sur la qualité des données.

· ***Les méthodologies :*** fournissent des directives pour choisir, a partir des techniques et des outils existants, la mesure de la qualité de données la plus efficace pour améliorer un système d'information spécifique.

· ***Les outils :*** qui sont nécessaires aux techniques et aux méthodologies, représentent des processus automatises avec une interface qui permettent a l'utilisateur l'exécution manuelle de certaines techniques pouvant être intégrées dans des structures (*Framework*).

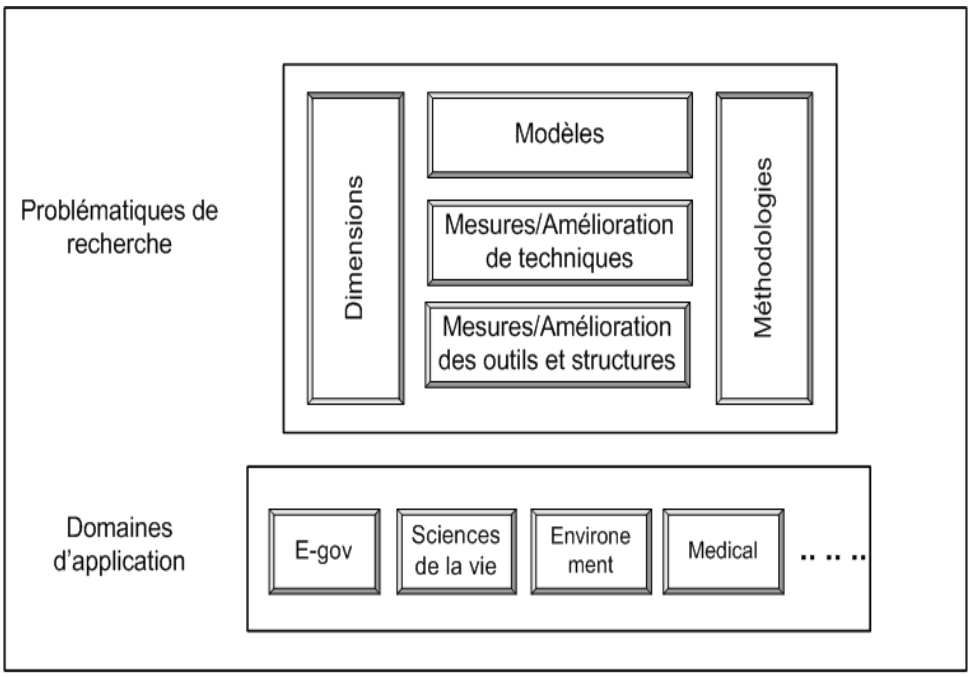
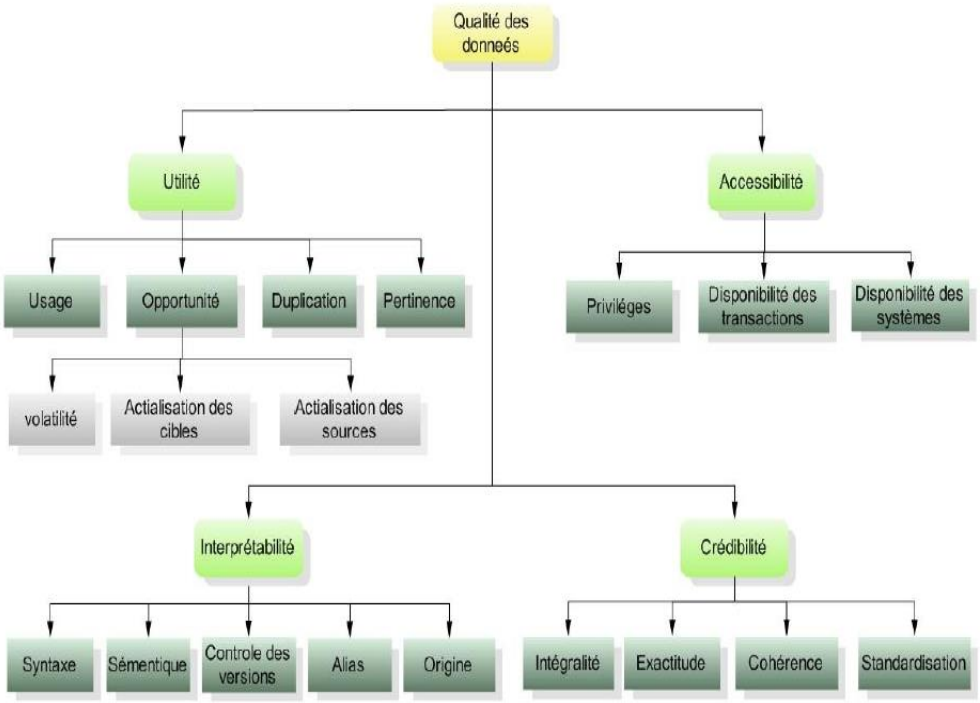


Figure 1.1 - Principales problématiques de la qualité des données

1.2.3 Dimensions de la qualité des données

Ont défini un ensemble de dimensions afin de mesurer de la qualité des données. Parmi ces dimensions, nous avons détaillé celles que nous allons utiliser dans notre processus afin de garantir aux utilisateurs des données ainsi que des décisions pertinentes. Les données doivent avoir la qualité nécessaire pour supporter le type d'utilisation. En d'autres termes, la demande de qualité est aussi importante sur les données nécessaires à l'évaluation d'un risque que sur celles utilisées dans une opération de marketing de masse.[4]



Figure(1-2): Les dimensions de la qualité des données

Plusieurs dimensions ont été identifiées pour la qualité de données, telles que l'exactitude (accuracy), la complétude (completeness), la consistance (consistency). Toute fois, il n'existe aucun accord général soit sur quel ensemble de dimensions définit la qualité des données, ou sur le sens exact de chacune des dimensions. Nous rappelons brièvement, ci-dessous, les définitions des dimensions essentielles [1]

Consistance

La consistance se réfère à la violation des règles sémantiques définies sur l'ensemble des données, Dans le modèle relationnel, les contraintes d'intégrités sont des instanciations de ces règles sémantiques.

Complétude

La capacité d'un système d'information à représenter chaque état significatif du monde réel représenté

Duplication

Les données sont répétées. L'entité est gérée par plusieurs systèmes d'informations sous des identifiants différents et donc sa vue n'est pas unifiée.

Exactitude (Accuracy)

Les données sont exactes lorsque les valeurs des données stockées dans les sources de données correspondent à celles du monde réel.

1.2.4 Les approches génériques

Ces approches peuvent être considérées comme théoriques car elles adoptent un modèle formel afin de justifier les dimensions, ou considérées aussi comme empiriques car elles construisent un ensemble de dimensions en accord avec des expérimentations, des entretiens et des questionnaires, ou intuitives car elles se basent sur un sens commun ou l'expérience pratique.[2]

1.2.4.1 Les approches théoriques

Les approches théoriques comme celle proposée par considère le système d'information comme la représentation du système du monde réel.

Le monde réel est représente si un système d'information existe, et est caractérise par le fait qu'en aucun cas deux états doivent être mis en correspondance au même état dans un système d'information. (Toute déviation dans cette représentation peut engendrer

Des déficiences, c'est-a-dire, des contraintes en termes de qualité. Les auteurs distinguent des déficiences, notamment liées a la conception et les opérations du Système. Plus spécifiquement, les représentations du système peuvent être *incomplètes*, *ambigües* ou avec des états *sans signification* et les opérations, peuvent être fausses. En accord avec ces déficiences, cette approche a de plus identifie certaines

Dimensions sur la qualité : l'exactitude, la fiabilité, la ponctualité, la complétude et la cohérence.

1.2.4.2 Les approches empiriques

Au niveau des approches empiriques, notamment dans les dimensions ont été choisies en interviewant les consommateurs de données Pour la sélection des dimensions, ils ont propose autour de 179 dimensions de la qualité des données, parmi lesquelles les auteurs en ont sélectionne quinze.[2]

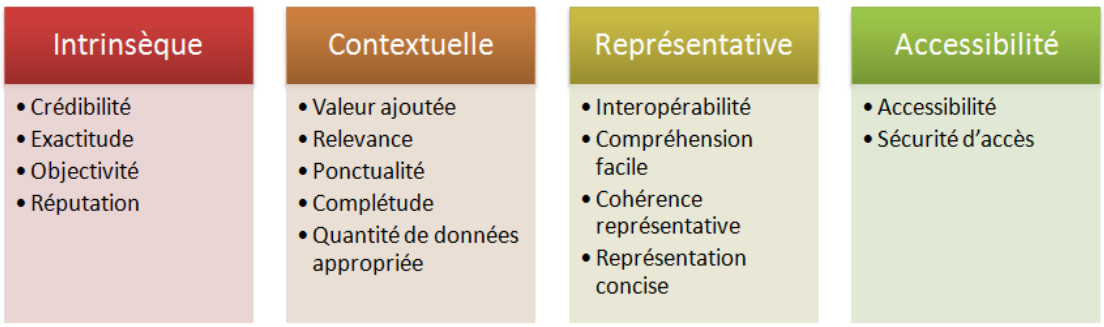


Figure 1.3 - Dimensions de la qualité de données proposées par l'approche empirique

Ces dimensions ont été classifiées selon quatre catégories (Figure 1.3) : qualité de données intrinsèques (la qualité concernant la donnée elle-même), qualité des données contextuelle (considère le contexte de l'utilisation de la donnée), qualité des données représentative (considère les aspects lies a la qualité de la représentation de la donnée) et qualité d'accessibilité des données (liée a l'accessibilité de données et aux propriétés non-fonctionnelles de l'accès aux données).[9]

1.2.4.3 Les approches intuitives

Concernant les approches intuitives, [Redman'97] a propose une classification des dimensions de la qualité selon le schéma conceptuel, les valeurs de données, et le format des données. Le schéma conceptuel

correspond aux dimensions de la qualité, qui sont liées aux données et sont indépendantes des représentations internes de la donnée.

Ces représentations sont gérées par les dimensions dédiées au format des données (Figure 1.4).

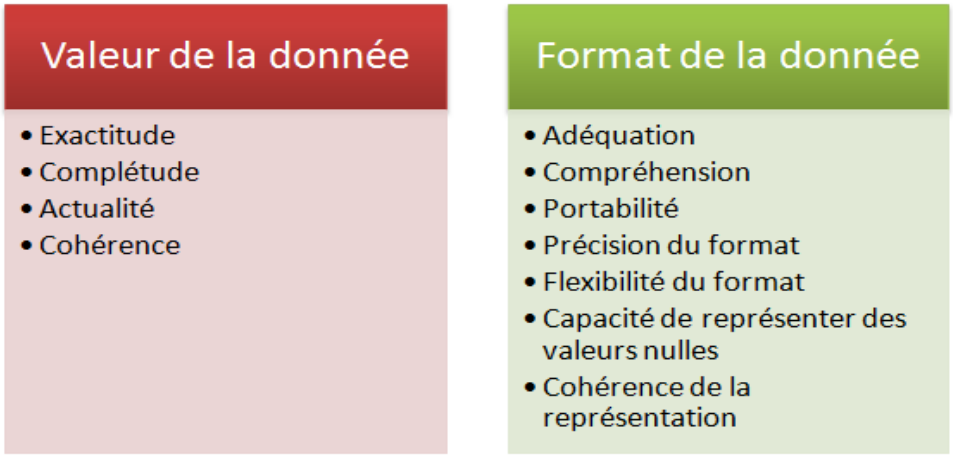


Figure 1.4 – Dimensions proposées dans une approche intuitive

La série d'approches proposées dans le domaine des systèmes d'information que nous avons présentée, nous amène a sélectionner les dimensions citées de façon récurrente (avec certaines divergences) par les travaux dans le domaine.[3] Ces dimensions récurrentes représentent le socle de dimensions utilisées pour caractériser la qualité des données Tableau (Figure 1.5).

Dimension	Description
Exactitude	Liée à « l'inexactitude », ce qui implique que l'état du monde réel est différente de ce qui est représenté.
Ponctualité	Définies le délai entre le changement de l'état du monde réel et la modification résultant de l'état du système d'information.
Complétude	L'habilité d'un système d'information à représenter chaque état du système du monde réel.
Cohérence	La cohérence apparait quand plus d'un état du système d'information correspond à un état du système du monde réel.
Compréhensibilité	Concerne la documentation et les métadonnées qui sont disponibles pour interpréter correctement la signification et les propriétés des sources de données.
Accessibilité	Mesure l'habilité que l'utilisateur accède aux données sans importer leur culture, leur statut/fonctions et les technologies disponibles.
Utilisabilité	Mesure l'effectivité, l'efficacité, la satisfaction avec laquelle les utilisateurs spécifiés perçoivent et utilisent les données.
Crédibilité	(D'une organisation) Mesure combien est fiable l'organisation qui fournis les sources de données.

Figure 1.5 – Dimensions récurrentes dans la qualité de données

2.2.5 Les approches spécifiques

Dans des domaines plus spécifiques (i.e. le web, la statistique, le domaine géographique), des dimensions plus adéquates ont été proposées afin de mieux décrier les spécificités de leurs domaines (*précision géométrique, cohérence temporelle, etc*.). D'autre part, le changement et la mise a jour des données ont conduit à des dimensions liées au facteur temps, c'est-a-dire, la qualité des données stables, variables à long terme, ou encore sur les données hautement ≪ changeables ≫. Les principales dimensions proposées dans ce contexte sont : *l'actualité*, *l'instabilité*, et *la ponctualité*

· ***L'actualité***, représente la rapidité avec laquelle la donnée est mise a jour ; elle peut être évaluée en considérant la dernière mise a jour (information gardée dans des

Métadonnées), c'est-a-dire, la dernière fois que la donnée a été actualisée.

· ***L'instabilité***, concerne la fréquence avec laquelle la donnée varie dans le temps, et peut être évaluée par la période du temps dont la donnée reste valide.

· ***La ponctualité***, exprime a quel point les données sont actuelles pour une tache, mais ceci n'implique pas seulement les données actuelles, mais aussi les périodes

De temps pour lesquelles une donnée a été planifiée. Pour mesurer la ponctualité, il s'agit de mesurer l'actualité de la mesure et contrôler si la donnée est disponible pour la période de temps planifie.

1.2.6 Les objectifs de qualité des données :

L'objectif principal de qualité des données est d'assurer l'exactitude, la pérennité, la pertinence et la consistance des données à travers une organisation ou à travers les différentes divisions d'une organisation et des lors assurer que les décisions prises le sont sur des informations consistantes et justes. [14]

En fonction des utilisations de la donnée, on détermine les critères qualité primordiaux à contrôler. Ensuite, on détermine les attributs de la donnée permettant de mesurer les critères qualité. Enfin, on spécifie le niveau minimal de qualité requise pour chaque critère retenu.

Les objectifs sont plus nombreux et plus diversifiés que dans le simple univers décisionnel. La granularité d'information et de suivi qualité doit répondre à tous les objectifs des processus consommateurs

1.3 Coût du non qualité

L'impact et donc le coût d'une donnée de mauvaise qualité n'est pas le même selon le type de population (dans un CRM (Customer Relationship Management), grand compte ou PME (petites et moyennes entreprises)) mais aussi selon l'utilisation qui en y faite (données bancaires, données médicales, données militaires sensibles ou données CRM). L'estimation des "coûts de la non-qualité" n'est pas aisée. Ajoutons que s'il est relativement aisé d'évaluer combien coûte la mise en œuvre d'une procédure d'amélioration, les bénéfices escomptés sont plus difficiles à chiffrer en raison des aspects non mesurables, mais néanmoins cruciaux, qui accompagnent l'amélioration de la qualité d'un système informatique, tels que la crédibilité ou la fiabilité de l'information. A titre indicatif, plusieurs études menées aux États-Unis dans des secteurs divers tels que banques, assurances ou agences de voyage font état d'un taux d'erreur de 5 % à 30 % dans les BSD (ce taux étant, par exemple, évalué sur la base du rapport entre le nombre d'enregistrements contenant au moins une erreur logique et le nombre total d'enregistrements d'une BD). En termes financiers, les coûts de la "non-qualité" sont évalués à une perte d'environ 5 à 10 % du revenu des entreprises examinées. Citons par exemple les coûts en contrôles, correction et maintenance de données de qualité douteuse, les coûts liés au traitement des plaintes des clients non satisfaits ou encore à la réparation des préjudices [9]

1.4 Conclusion :

Dans ce chapitre, nous avons présenté un état de l'art sur la qualité des données afin d'aboutir à l'identification des limitations des différents

travaux et approches dans les différents domaines nécessaires à l'amélioration de la qualité des données.

L'état de l'art que nous avons présenté dans ce chapitre sur la qualité des données et les entrepôts de données et de leur qualité est très utile dans notre travail, dans l'amélioration de leur qualité ainsi que le nettoyage des données.

Chapitre2

Couplage d'enregistrement

2.1 Introduction

Le couplage d'enregistrements est le processus d'identification des enregistrements faire référence aux mêmes entités du monde réel dans des situations où ne sont pas disponibles. Les enregistrements sont liés à la base de la similitude entre attributs communs, chaque paire étant classée comme « lien » ou « non-lien » en fonction de leur similitude.[5]

Dans ce chapitre, nous présentons couplage d'enregistrements, nous pouvons classer ces travaux suivant deux approches : les méthodes déterministes et les méthodes probabilistes, comparons plusieurs algorithmes de liaison : blocage traditionnel, incomplet méthode de recherche de similarité, LSH-MinHash, et une méthode complète, M-tree.Nous détaillons chacune de ces méthodes dans la suite

2.2 Le couplage d'enregistrements

2.2.1Définition

Couplage d'enregistrements (en anglais, record linkage), également connu sous le nom de résolution d'entité, de correspondance de données et de doublon détection, est le processus d'identification et de mise en correspondance d'enregistrements mêmes entités du monde réel dans ou entre les ensembles de données.[6]

Les entités à relier sont souvent des personnes (comme des patients hospitalisés ou des clients dans des ensembles de données d'entreprise), le couplage d'enregistrements peut également être appliqué pour relier

des produits de consommation ou des références bibliographiques. Enregistrements [7].

Le couplage d'enregistrements est souvent contesté par le manque d'entité unique identifiants (clés) dans les ensembles de données à lier, ce qui empêche l'utilisation d'une base de données joindre. Au lieu de cela, le couplage des enregistrements nécessite la comparaison des les attributs disponibles dans les jeux de données, par exemple noms, adresses et dates de naissance des personnes.

Attribut	Dataset-1	Dataset-2
Prénom	Alice	Alicia
Nom	Smith	Smith
Date de naissance	19950821-1320	199508211320
Num de téléphone	265-5984156	151-0484631

FIGURE 2.1 contient les enregistrements de deux personnes. Il est représentatif du couplage d'enregistrements problème car il visualise que la même personne peut avoir différents attributs attribués dans différents jeux de données et correspondent toujours à la même entité du monde réel.

En utilisant des approches de couplage d'enregistrements, il est possible d'identifier les relations et de catégoriser les données selon des entités du monde réel. Le tableau de FIGURE 2.1 montre que le

même attribut personne peut être représenté sous différentes formes dans différentes sources de données. En utilisant des fonctions de comparaison sur les données numériques et textuelles dans un processus de liaison, puis en utilisant des algorithmes de classification nous pouvons déterminer si une paire est compatible ou non. Avec le nombre croissant de produits distribués et hétérogènes, des solutions efficaces de couplage d'enregistrements sont très utiles pour trouver un vue unifiée des données.

2.2.2 Méthodes déterministes

Les méthodes déterministes, se basent sur des règles définies par des experts déterminant les conditions de couplage d'une paire d'enregistrements. Ces règles sont généralement dépendantes d'un ensemble de champs pertinents (dits variables de couplage). Si les attributs d'une paire donnée d'enregistrements correspondant aux champs pertinents coïncident, alors cette paire est couplée. Des poids (par exemple, la fréquence du champ) peuvent être attribués à ces champs. Ainsi, si la somme pondérée du nombre d'attributs qui coïncident dépasse un seuil, alors le couplage est retenu. Souvent, les comparaisons sont de type exact et ne tolèrent pas les erreurs de saisie. [7] Ces techniques sont coûteuses en temps car elles nécessitent une implication importante de l'utilisateur pour permettre la génération de transformations spécifiques au domaine et aux données. De plus, elles sont trop rigides (car elles sont dépendantes de la base de données) pour corriger les erreurs évoquées dans l'introduction de cette partie

2.2.3 Méthodes probabilistes

Les méthodes probabilistes consistent à utiliser les statistiques sur les propriétés des variables en commun entre paires d'enregistrements pour calculer la probabilité qu'ils représentent la même entité.[7]

Elles peuvent être non supervisées ou supervisées.

2.2.3.1Méthodes non supervisées

Ces méthodes sont intéressantes quand il n'existe pas de données annotées. Un modèle théorique de couplage d'enregistrements qui offre des méthodes statistiques pour estimer les paramètres de couplage et les taux d'erreurs a été proposé par Fellegi et Sunter dans [10]. Ce modèle est décrit ci-après.

Modèle de Fellegi et Sunter

Définition : Le modèle définit les quantités suivantes pour fonctionner :

— M-probabilité : c'est la probabilité de couplage (matching) estimée pour qu'une paire d'attributs quelconque d'un champ donné soit similaire sachant que la paire d'enregistrements correspondante est un vrai couplage (i.e. les deux enregistrements correspondent en réalité);

— U-probabilité : c'est la probabilité de non couplage (unmatching) estimée pour qu'une paire d'attributs quelconque d'un champ donné soit similaire sachant que la paire d'enregistrements correspondante est un faux couplage, i.e. la probabilité que les deux enregistrements se couplent par hasard;

— ratio de couplage d'attributs : le ratio de couplage d'une paire d'attributs correspondants à un champ ayant une probabilité de couplage M et une probabilité de non couplage U est calculé par :

— pour un accord sur l'attribut : log(M/ U);

— pour un désaccord sur l'attribut : log(1−M/ 1−U);

— ratio de couplage d'enregistrements R : le ratio de couplage d'une paire d'enregistrements est calculé par la somme des ratios de couplage des différentes paires d'attributs qui le composent; [11]

— le couplage d'enregistrements se fait par rapport aux seuils Tλ et Tμ comme suit :

— si R < Tλ alors non couplage (rejet);

— si Tμ < R < Tλ alors possible couplage (indécis);

— si R > Tμ alors couplage (acceptation).

Problèmes posés. Les problèmes posés par ce modèle se traduisent par 3 questions :

— Comment définir M et U ?

— Comment fixer les seuils Tλ et Tμ ?

— Comment déterminer les mesures de similarité entre champs ?

Interprétation. Pour l'évaluation, nous pouvons définir les vrais et les faux positifs (resp. négatifs) à l'aide de ces considérations de couplage (resp. non couplage) et de liaison (resp. non liaison) (voir Tableau). En effet, si nous avons décidé de coupler (resp. ne pas coupler) une paire d'enregistrements alors si cette paire est liée (resp. n'est pas liée) en

réalité, alors il s'agit d'un vrai positif (resp. vrai négatif) sinon il s'agit d'un faux positif (resp. faux négatif).[13]

	Couplage	Non couplage
Liaison	vrai positif (TP)	faux positif (FP)
Non liaison	faux négatif (FN)	vrai négatif (TN)

FIGURE 2.2 – Classification des résultats de couplage

Plusieurs méthodes se sont basées sur le modèle de Fellegi et Sunter pour le couplage d'enregistrements. A titre d'exemple, Winkler a détaillé une technique qui se base sur l'algorithme "espérance-maximisation" pour estimer les paramètres du modèle et optimiser les règles de couplage. Dans le cas des bases de données volumineuses (par exemple, des dizaines de milliers d'enregistrements), le couplage d'enregistrements consiste à comparer tous les enregistrements deux à deux en calculant le produit cartésien, ce qui n'est pas efficace en temps et en mémoire. Pour résoudre ce problème, les auteurs dans proposent une amélioration du modèle de Fellegi et Sunter consistant à séparer la base de données en des groupements d'enregistrements (blocs ou fenêtres), en se basant sur de simples heuristiques qui aident à éliminer les paires d'enregistrements clairement non liés.**[11]**

2.2.3.2 Méthodes supervisées

Dans ce type de méthodes, les auteurs traitent le problème de couplage d'enregistrements se référant à la même entité comme un problème de classification supervisée.[7] En effet, ils proposent de représenter chaque paire d'enregistrements à l'aide d'un vecteur de caractéristiques décrivant les similarités entre les paires d'attributs. Ces caractéristiques peuvent être binaires (par exemple, les attributs "nom"

correspondent), discrètes (par exemple, les n-premiers caractères du “prénom” correspondent) ou continues (par exemple, la mesure de Levenshtein entre les prénoms). Ainsi, la comparaison entre paires d’enregistrements conduit à les ranger dans les 3 classes : couplage, non couplage ou couplage possible. Les classifier utilisés déterminent, à partir des données annotées, les conditions de couplage entre les enregistrements, à savoir la détermination des champs pertinents et les seuils de décision pour la similarité entre les attributs. Nous pouvons citer à titre d’exemple les travaux dans [Bilenko03b] qui proposent une méthode basée sur deux niveaux d’apprentissage supervisé employant un SVM. Le premier niveau représente la comparaison de paires d’attributs en utilisant un apprentissage avec la mesure de Levenshtein (similarité au niveau attribut) et le deuxième niveau représente la comparaison des paires d’enregistrements en utilisant un apprentissage sur la similarité (similarité au niveau enregistrement).**[7]**

Les auteurs dans [Tejada02] utilisent un apprentissage actif pour sélectionner les paires d’enregistrements les plus informatives. Ainsi, l’utilisateur est sollicité pour annoter ces paires d’enregistrements informatives comme paires liées ou non liées afin d’entraîner les classifier. Ces derniers génèrent des règles de classification qui croisent des attributs et des mesures de similarité entre ces attributs. Les auteurs dans [Churches02] proposent de normaliser certains attributs (ceux qui désignent des noms et des adresses postales) pour effectuer convenablement la comparaison de ces attributs dans le processus de couplage d’enregistrements. Cette normalisation est effectuée en se basant sur un résultat de segmentation des attributs (par exemple un attribut adresse est segmenté en : numéro de voie, type de voie, nom de voie). Pour la segmentation, un modèle de Markov caché est utilisé avec

comme observations, les mots de l'attribut, et comme états cachés, les segments. Cette méthode nécessite suffisamment de données annotées sous forme de chaînes de caractères segmentées pour l'apprentissage

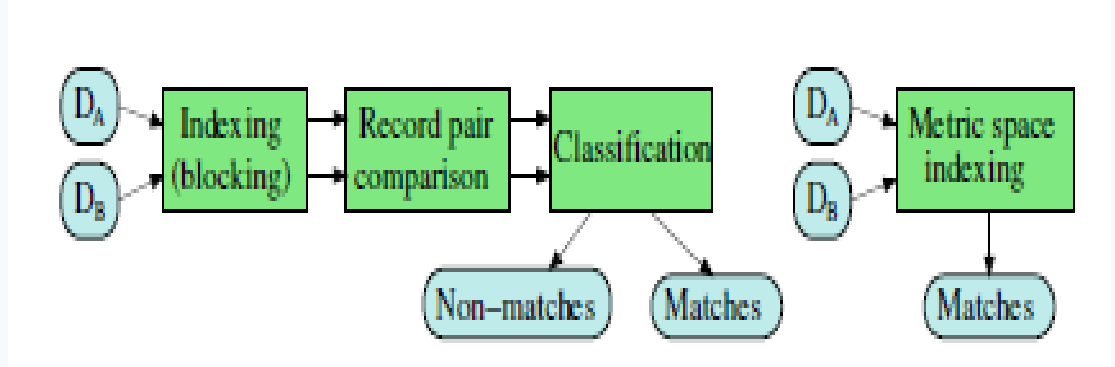


FIGURE 2.3: Aperçu des étapes du processus de couplage d'enregistrements traditionnel (côté gauche) et notre approche basée sur l'indexation de l'espace métrique proposée (côté droit), comme décrit dans Sect. 1, où les enregistrements de deux ensembles de données, DA et DB, sont liés.

2.2.4 Approche

Nous abordons le problème de liaison général suivant :

pour deux ensembles de données DA et DB, nous souhaitons trouver, pour chaque enregistrement dans DA, tous les enregistrements dans DB qui correspondent par rapport à un certain seuil de distance d (c'est-à-dire avoir une distance de d ou moins) Nous comparons plusieurs algorithmes de liaison: blocage traditionnel, incomplet méthode de recherche de similarité, LSH-MinHash, et une méthode complète, M-tree. Nous utiliser également une technique de force brute complète comme base, bien que cela ne puisse être appliqué à notre plus petit ensemble de données.

Toutes les expériences ont un certain nombre de paramètres pour configurer l'espace de recherche et le comportement de l'algorithme, y fonction de distance et le seuil, d, spécifiant la distance maximale pour deux les enregistrements à classer comme un lien (c'est-à-dire faisant référence à la même entité). [16]

Nous concentrons sur une seule fonction de distance dans ces expériences.

2.2.4.1Brute force: Chaque enregistrement dans DA est comparé à chaque enregistrement dans DB. Chaque la paire est classée comme un lien si la distance entre les enregistrements est inférieure ou égale au seuil d. Cela $O(\left|DA\right|.\left|DB\right|)$[6]

2.2.4.2Blocage traditionnel: les paramètres sont l'ensemble des clés de blocage et (facultativement) les codages phonétiques appliqués à chaque attribut. Ceux-ci sont sélectionnés comme décrit en exploitant la connaissance du domaine et des données, et choisis dans le but de donner les meilleurs résultats possibles. Chaque enregistrement dans DA est placé dans le bloc approprié en fonction de sa valeur de clé de blocage.

Le l'algorithme itère ensuite les enregistrements DB dans la base de données, et pour chacun le compare avec chacun des enregistrements de DA dans le bloc avec la même valeur de clé de blocage.[17]

Dataset name(s)	Records in Dataset DA	Records in dataset DB	Number of true matching pairs	Entities linked
Cora	1,295	1,295	17,184	Publication_Publication
Isle of Skye	17,612	12,284	2,900	Birth_Death
Kilmarnock	38,430	23,714	8,300	Birth_Death

FIGURE 2.4 :Tableau Caractéristiques des ensembles de données utilisés dans les expériences.

2.2.4.3 LSH-MinHash: Les paramètres de LSH-Minhash sont (shingle size (lss),band size (lbs) et number of bands (lnb)). Tout d'abord, les attributs de chaque enregistrement dans Les DA sont concaténés, et le résultat

shingle se transforme en un ensemble de n-grams avec n= lss.

Ensuite, un ensemble de fonctions de hachage générées de manière déterministe est appliqué à chaque n-gram dans l'ensemble et le plus petit résultat (le MinHash) de chaque application de hachage est ajouté à une signature pour l'enregistrement. Le nombre de hachages utilisés, et donc la taille de la signature est fixée à lnb×lns.

Enfin, la signature est divisée en lnb les bandes et les valeurs de chaque bande sont hachées à nouveau pour créer un certain nombre de clés. L'enregistrement d'origine est ajouté à une carte associée à chacune des clés.

Pour effectuer le couplage, l'algorithme parcourt les enregistrements de la base de données. Chaque enregistrement est hachée comme décrit ci-dessus, pour obtenir un jeu de clés. Chaque clé est recherchée dans la structure de données et les enregistrements associés de DA ajoutés au jeu de résultats.[16]

Enfin, l'enregistrement de DB est comparé à son tour à chaque enregistrement du résultat ensemble, la paire étant classée comme un lien ou un non-lien en fonction de leur distance.[17]

2.2.4.4M-tree: l'algorithme de liaison n'a pas de paramètres supplémentaires. Comme avec LSH-MinHash, chaque enregistrement dans DA est inséré dans un M-tree. Pour effectuer la liaison, l'algorithme itère sur chaque enregistrement b € DB. Une opération de recherche par plage (b, d) est effectuée sur l'arbre M, en passant le seuil de distance d comme deuxième paramètre Tous les enregistrements retournés sont directement classés en tant que liens.[6]

2.3 Conclusion

Dans ce chapitre nous présentons les méthodes de couplages d'enregistrement. Ces méthodes se basent sur la comparaissent d'attributs. Il est alors intéressant d'étudier les manières de combinaison de masures de similarité pour le couplage d'enregistrement.

CHAPITRE 3

LA CONCEPTION

3.1 Introduction

Dans ce chapitre nous présentons les méthodes de résolution d'entités qui se base sur la comparaison d'attributs pour le couplage d'enregistrements se référant à une même entité. Pour la comparaison d'attributs, nous étudions différentes manières de combinaisons entre les mesures de similarité.

3.2 Méthodes de résolution d'entités

La résolution d'entités se base principalement sur le couplage d'enregistrements se référant à une même entité. En plus, une phase préliminaire qui permet de réduire le nombre de comparaisons et une phase finale de structuration des enregistrements couplés en entités, sont intégrées. [13]

3.2.1 Pré-couplage

Le pré-couplage consiste à segmenter la base de données en des blocs d'enregistrements à l'aide d'un regroupement des enregistrements qui ont des chances de représenter la même entité dans un même bloc. Ainsi, seuls les enregistrements d'un même bloc sont comparés entre eux dans la phase de couplage. Ce regroupement est réalisé à l'aide de clés de regroupement (par exemple, un champ donné, une combinaison de champs, etc.). Les enregistrements dont les attributs, qui correspondent à ces clés, coïncident sont regroupés. Comme exemples de clés, nous pouvons considérer les n-premiers caractères du champ "nom" d'une entreprise ou les n-premiers termes du champ "titre" d'une référence bibliographique.[7]

3.2.2 Couplage d'enregistrements

Le couplage d'enregistrements se base sur la comparaison de leurs paires d'attributs. Un problème important consiste à sélectionner la mesure de similarité adéquate pour cette comparaison. Une fois la mesure sélectionnée, une méthode de prise de décision de couplage, qui tient compte des résultats de comparaison d'attributs, doit être utilisée.

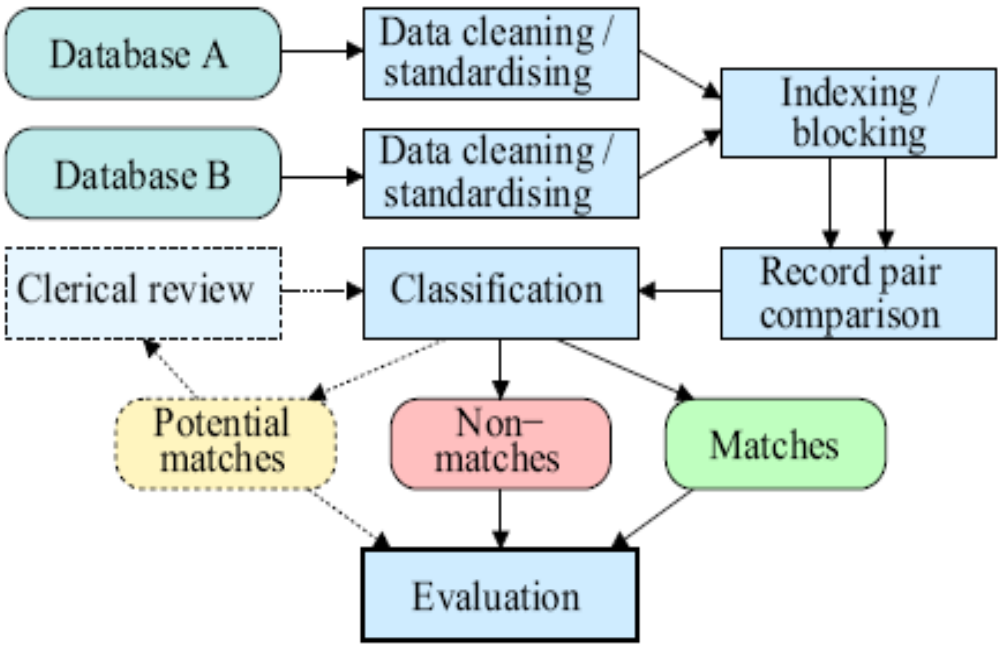


FIGURE 3.1: Procédure générale Couplage d'enregistrements

La tâche principale du nettoyage et de la normalisation des données est la conversion des données d'entrée brutes en données bien définies et cohérentes formulaires, ainsi que la résolution des incohérences dans la manière dont l'information est représenté et codé.

Si deux ensembles de données, A et B, doivent être liés, potentiellement chaque enregistrement A doit être comparé à tous les enregistrements de B. [14]

Le nombre d'enregistrements possibles, les comparaisons de paires sont donc égales au produit de la taille des deux ensembles de données, $|A| \times |B|$. De même, lors de la déduplication d'un ensemble de données, A, le nombre deles paires d'enregistrements sont $|A| \times (|A| - 1) / 2$.

Le goulot d'étranglement des performances dans une donnéele système de liaison ou de déduplication est généralement la comparaison détaillée coûteuse

de champs (ou attributs) entre des paires d'enregistrements, ce qui rend impossible comparez toutes les paires lorsque les ensembles de données sont volumineux. Par exemple, lier deux données

ensembles avec 100; 000 enregistrements chacun donneraient 1010 (dix milliards) d'enregistrements comparaisons. D'autre part, le nombre maximum de correspondances vraies quisont possibles correspond au nombre d'enregistrements dans le plus petit jeu de données(en supposant qu'un enregistrement dans A ne peut être lié qu'à un maximum d'un enregistrement dans B, et vice versa). Par conséquent, le nombre de correspondances potentielles augmente linéairement lors de la liaison d'ensembles de données plus volumineux, alors que les efforts de calcul augmentent Quadratiquement. La situation est la même pour la déduplication, où le nombred'enregistrements en double est toujours inférieur au nombre d'enregistrements dans un ensemble de données. Pour réduire le grand nombre de comparaisons possibles de paires

d'enregistrements, les techniques de liaison de données utilisent le blocage, c'est-à-dire qu'elles utilisent un ou une combinaison

d'attributs d'enregistrement (appelés la variable de blocage) pour fractionner les données ensembles en blocs.

Tous les enregistrements ayant la même valeur dans la variable de blocage seront placés dans le même bloc, et seuls les enregistrements d'un bloc seront comparés.

Cette technique devient problématique si une valeur de la variable de blocage est mal enregistrée, car un enregistrement potentiellement correspondant peut être inséré dans un bloc différent, interdisant la possibilité d'un match. Pour surmonter ce problème, plusieurs passes (itérations) avec différentes variables de blocage sont normalement effectué.

3.2.2.1Couplage d'enregistrements probabilistes

parfois appelé appariement flou (aussi fusion probabiliste ou fusion floue dans le contexte de la fusion des bases de données), adopte une approche différente du problème de couplage d'enregistrements en prenant en compte un éventail plus large d'identificateurs potentiels, les poids de calcul pour chaque identifiant en fonction sur sa capacité estimée d'identifier correctement un match ou d' un non-match, et l' utilisation de ces poids pour calculer la probabilité que deux enregistrements données se rapportent à la même entité. Les paires d'enregistrements avec des probabilités au- dessus d' un certain seuil sont considérés comme des matchs, tandis que les paires avec probabilités ci dessous un autre seuil sont considérés comme non-matches; paires qui se situent entre ces deux seuils sont considérés comme des

« correspondances possibles » et peuvent être traitées en conséquence (par exemple, l' homme examinés, liés ou non liés, selon les besoins).[13] Alors que le couplage d'enregistrements déterministe nécessite une série de règles potentiellement complexes à programmer à l' avance, les méthodes de couplage d'enregistrements probabiliste peuvent être « formés » pour bien performer avec une intervention beaucoup moins humaine.

De nombreux algorithmes de couplage d'enregistrements probabiliste affecter de correspondance / non-correspondance des poids à des identificateurs au moyen de deux probabilités appelées u et m .
La u probabilité est la probabilité qu'un identificateur de deux non-correspondance des enregistrements sera d' accord par pur hasard. Par exemple, la u probabilité pour le mois de naissance (où il y a douze valeurs qui sont à peu près uniformément répartis) est $1/12 \approx 0{,}083$; identificateurs avec des valeurs qui ne sont pas uniformément répartis auront différentes u probabilités pour des valeurs différentes (y compris éventuellement les valeurs manquantes). Le m probabilités est la probabilité qu'un identificateur correspondant à des paires sera d' accord Cette valeur serait de 1,0 dans le cas des données parfaites, mais étant donné que cela est rarement (voire jamais) vrai, celle - ci peut être estimée. Cette estimation peut être faite sur la base de la connaissance préalable des ensembles de données, en identifiant manuellement un grand nombre de correspondants et les paires ne correspondent pas à « train » de l'algorithme de couplage probabiliste des enregistrements, ou par itérativement l' exécution de l'algorithme pour obtenir des estimations plus étroites de la m probabilité. Si une valeur de 0,95 devait être estimée pour la m probabilité, puis le match de / poids non-match pour l'identificateur de mois de naissance seraient :

Résultat	Proportion de liens	Proportion des non-liens	rapport de fréquence	Poids
Rencontre (match)	m = 0,95	u 0,083 ≈	$\frac{m}{u}$≈ 11,4	ln ($\frac{m}{u}$) / ln (2) ≈ 3,51
Non-correspondance (non_match)	1- m = 0,05	1- u 0,917 ≈	$\frac{1-m}{1-u}$≈ 0,0545	ln ($\frac{1-m}{1-u}$) / ln (2) ≈ -4,20

FIGURE 3.2 : *tableau de Couplage d'enregistrements probabilistes*

Les mêmes calculs seraient effectués pour tous les autres identifiants à l'étude pour trouver leur match / poids non-match. Ensuite, chaque identifiant d'une fiche serait comparé avec l'identificateur correspondant d' une autre fiche pour calculer le poids total de la paire: la *partie* du poids est ajoutée au total cumulé à chaque fois une paire d'identificateurs d' accord, tandis que le *non-correspondance* poids est ajouté (les diminutions totales de fonctionnement) toutes les fois que la paire d' éléments d'identification est en désaccord. Le poids total obtenu est ensuite comparé à des seuils mentionnés ci - dessus pour déterminer si la paire doit être liée, non-lié, ou mis de côté pour une considération particulière (par exemple , la validation manuelle).[15]

Proba(name, surname-sac- sec-id) = → m=0.95
U=0.01

Weight= log2($\frac{m}{u}$) = → 6.57 match
log2 ($\frac{1-m}{1-u}$)=-4.31 un match

proba(strect-number-addr1-addr2-sumb-postecod)
m=0.7
u=0.01

weight= log2 ($\frac{m}{u}$) = 6.13 matche
weight non match = log2($\frac{1-m}{1-u}$)=-1.72

(date de naissance- age) = m0.9 et u0.01
Weight = → log2($\frac{m}{u}$)= 6.49 match
log2 ($\frac{1-m}{1-u}$)= -3.31 non match

- name r1 =name r2 ⇨ w =log2 ($\frac{0.95}{0.01}$)= 6.57
- surame r1 = surname r2 ⇨ w = log2($\frac{0.95}{0.01}$) = 6.57
- strect number r2 = strectnember r2 ⇨ w= log2 (0.7/0.01)=6.13
- addr r1≠ addr r2 ⇨ w =log2($\frac{1-0.7}{1-0.01}$) = -1.72
- addr r1= vide alors ⇨ w= 0
- suburb r1= suburb r2 ⇨ w= $\log2\left(\frac{0.7}{0.01}\right) = 6.13$
- post code r1 = poste code r2 ⇨ w = log2($\frac{0.7}{0.01}$) = 6.13
- state r1 = state r2 ⇨ w = $\log2\left(\frac{0.7}{0.01}\right) = 6.13$
- date B r1 = date B r2 ⇨ w= $\log2\left(\frac{0.9}{0.01}\right) = 6.49$
- age r1, r2 = vide alors ⇨ w = 0

- phone_number r2 = phone_number r2 ⟹
  w=log2 $\left(\frac{0.7}{0.01}\right) = 6.13$
- sec_d r2 = sec_d r2 ⟹ w= log2$(\frac{0.95}{0.01})$ = 6.57
- Le total = 55.13

3.2.2.2 Comparaison d'enregistrements le (modèle de Fellegi et Sunter) :

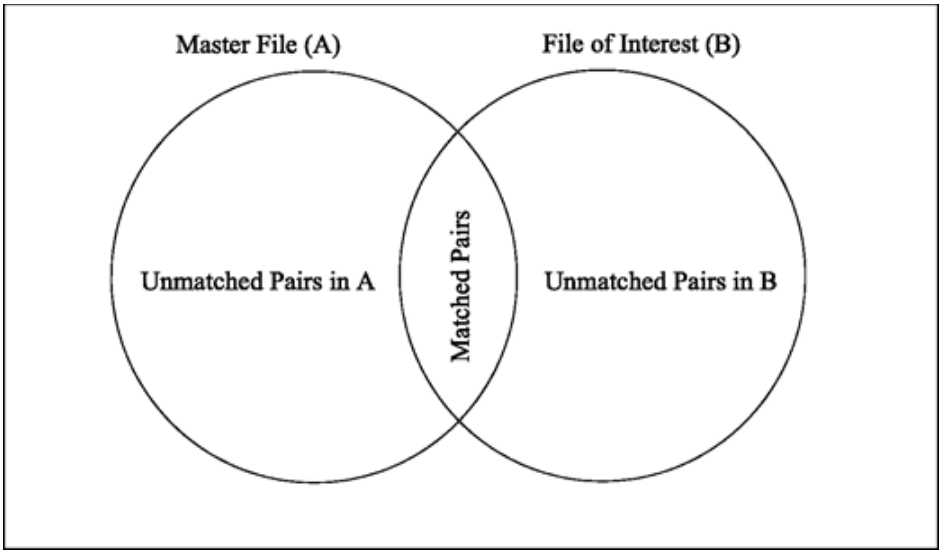


FIGURE 3.3 **:** Partitionnement de deux fichiers en enregistrements correspondants et sans correspondance.

Elle est faite par rapport à un score probabiliste calculé en fonction de la similarité [11]

. Méthode proposée entre les paires d'attributs et les probabilités de couplage estimées pour les champs de la base de données. Cette comparaison est décrite par **l'Algorithme** où nous notons:

— E est la base de données définie par n enregistrements {ri} et m champs {cj}; chaque enregistrement ri est composé de m attributs, où l'attribut ri .cj correspond au champ cj ;

— q est le nombre de blocs {Bk} regroupant les enregistrements de E ;
— |Bk| est la taille (i.e. nombre d'enregistrements) du bloc Bk ;

— sim(r1.c, r2.c) est la fonction Maximum des mesures entre deux attributs r1.c et r2.c ;

— M(c) et U(c) représentent respectivement les probabilités M et U estimées pour un champ c ;

— Ratio est le ratio de couplage calculé pour chaque paire d'enregistrements ;

— Tλ et Tμ représentent respectivement les seuils pour la mesure de similarité entre attributs sim et le ratio de couplage Ratio. Ces seuils sont fixés expérimentalement à l'aide d'une base de validation.

L'algorithme consiste à comparer deux à deux les paires d'enregistrements dans chaque bloc. Cette comparaison se base sur un ratio de couplage Ratio défini en fonction de la similarité entre les paires d'attributs d'un même champ (si sim dépasse le seuil Tλ alors les attributs sont considérés comme similaires) et les probabilités M et U.

Dans le cas où le ratio de couplage dépasse le seuil Tμ, les enregistrements sont considérés se référant à la même entité et donc liés.

Dans une application avec deux fichiers, A et B, désignent les lignes (enregistrements) par au fichier A et dans le fichier B. Affecter caractéristiques à chaque enregistrement. L'ensemble des dossiers qui représentent des entités identiques est défini par α(a)β(b) K

M = {(a, b); a =b; a € A; b € B}

et le complément de jeu, à savoir mettre en représentant différentes entités est définie comme MU

U= {(a,b); a≠b; a € A; b € B}

Un vecteur, est défini, qui contient des accords et des désaccords codées sur chaque caractéristique : ϒ

3.3Algorithme de Couplage d'enregistrements:

Algorithme 1 : Couplage d'enregistrements

```
1algorithme: {
2    input
3     data base A, data base B
4     tab1= Read (Data1)
5      tab2= Read (Data2)
6       Bls= Blocking (tab1, tab2 );
7       Record pair comparaison (Bls);
8 weight record pair comparison     (pairs)
9   {
10          for (i=0 :pair.size )  {
11for(j=0 : attribut) {
12if(r1.att = r2.att){W=log (M/p)}
13                  else { W=log (1-m/u)}
14  }
15            }
```

READ

```
1   Read (data) {
2     tab [record] records
3   for All Feilds in data {
4         record A;
5  for All attribut in feilds {
6  R.addattribut (feildsi)
7  }
8  add R to tab
9   }
10  return tab;
11  }
```

Block

```
1   Blocking (tab1,tab2) {
2 metre les records qui ont meme date de naissance dans
le          meme block
3      return tab [blocking]
4          }
5      record par comparison (blocks) {
6        for ( All Block in blocks ) {
7          for (i=0 : nb record in Block) {
8            for(j=i+1: nb record in block ) {
9                weight record pair (record i, record j) }
10           }
11         }
12       }
```

3.4 Conclusion :

Dans ce chapitre, nous avons étudié l'étape de résolution d'entités qui a comme but de représenter les enregistrements d'une base de données dans un modèle entité factorisé. Cette étape se base sur le Couplage d'enregistrements qui inclue la comparaison de paires d'attributs et la comparaison de paires d'enregistrements. Pour ce faire, nous avons choisir couplage probabiliste, il peut être décomposé en un nombre relativement restreint d'opérations simples de manipulation de données avec relativement peu de connaissances statistiques.

CHAPITRE 4

IMPLÉMENTATION

4.1 Introduction

Après avoir présenté les méthodes de couplage d'enregistrements et les modèles utilisés dans le chapitre précédent, nous passons dans ce chapitre aux concepts techniques liés à l'implantation. Nous commençons par présente l'environnement opérationnel les Data set puis nous décrivons l'environnement de l'implémentation (langage de programmation utilise et l'environnement de développement) ensuite nous présentons notre application où nous évaluons Algorithme de couplage d'enregistrements sur des bases de données réelles. Nous terminons par les tests effectues et les résultats obtenus.

4.2 Data Set :

Similaire a notre problème. Le data set se compose de XX attribut indique l'ID de la personne ; le deuxième attribut indique sa position dans l'axe des (X) et le troisième attribut dans l'axe des (Y), le quatrième n'est pas utilise pour notre problème le cinquième et le sixième indique respectivement le temps de départ et d'arrivée de chaque individus la première ligne de notre data set nous fournit des informations sur la destination

4.3 Environnement de l'implémentation

4.3.1 Le langage de programmation

Nous avons choisi le langage JAVA, ce choix se justifie par :

- JAVA est un langage multiplateformes qui permet aux concepteurs, selon le principe: «write once, run everywhere », d'écrire un

code capable de fonctionner dans tous les environnements (quelque soit le système d'exploitation).

- Java assure une totale indépendance des applications vis-à-vis de l'environnement d'exécution, c'est-à-dire que toute machine supportant Java est en mesure d'exécuter un programme sans aucune adaptation (ni recompilation, ni paramétrage des variables d'environnement).

- JAVA est un langage orienté objets, simple qui réduit le risque d'erreurs et d'incohérence.

- JAVA est doté d'une riche bibliothèque de classe couvre de nombreux domaines (gestion de collection, accès aux bases de données, interface utilisateur graphique, accès aux fichiers et aux réseaux, utilisation d'objets distribués, XML….) sans compter toutes les extensions qui s'intègrent facilement à java.

- Un accès simplifié aux bases de données, soit à travers la passerelle JDBC-ODBC ou à travers un pilote JDBC spécifique au SGBD.

- Portabilité de l'exécutable : un programme java, une fois écrit et compilé, peut être exécuté sans modification sur tout système qui prend en charge java.

- Le développement avec java est gratuit.

4.3.2 L'environnement de développement

4.3.2.1 NetBeans :

-NetBeans est un projet open source ayant un succès et une base d'utilisateur très large. Sun Microsystems a fondé le projet open source NetBeans en Juin 2000 et continue d'être le sponsor principal du projet.

-Conçu en Java, NetBeans est disponible sous Windows, Linux, Solaris (sur x86 et SPARC), Mac OS X [18].

Aujourd'hui, deux projets existent: L'EDI NetBeans et la Plateforme NetBeans.

- L'EDI NetBeans (Environnement de Développement Intégré) : est un environnementϖ de développement - un outil pour les programmeurs pour écrire, compiler, déboguer et déployer des programmes. Il est écrit en Java - mais peut supporter n'importe quel langage de programmation. Il y a également un grand nombre de modules pour étendre l'EDI NetBeans [18].
- La Plateforme NetBeans : c'est également une plateforme. Il vous est possible de créerϖ votre propre application Awt ou Swing, basée sur la plateforme NetBeans.

4.3.2.2JavaFX :

- JavaFX est une famille de produits et de technologies de Sun Microsystems qui appartient à Oracle, qui base sur la machine virtuel Java pour fonctionner donc la communication avec des applications java standard est très simple. Une application JavaFX a accès à toutes les classes fournies par la machine virtuelle java [19].

-Les produits JavaFX ont pour but de créer des applications internet riches (RIA) et Facilite le développent avec images, graphiques, audio et vidéo. Actuellement JavaFX est constitué de JavaFX Script et de JavaFX Mobile, bien que d'autres produits soient prévus [20].

Les avantages :

- Basé sur Java (Java SE et ME).

- Utilisable sur tous les écrans : navigateurs, mobile, TV, etc.
- Open Source.
- Déploiement sur navigateur et ordinateur de bureau "Desktop" sans modification.
- Collaboration designers et développeurs. Possibilité d'intégrer des codes en Java et JavaFX
- Moins de code pour générer une interface et des composants graphiques(NSY).

4.4 Présentation de L'application

L'application que nous avons développé, en se basent sur les solutions proposé lors de la conception présenté dans le chapitre précédent, 1 -Tout d'abord, nous montrons l'interface principale de notre application Dataset sur laquelle nous voulons travailler

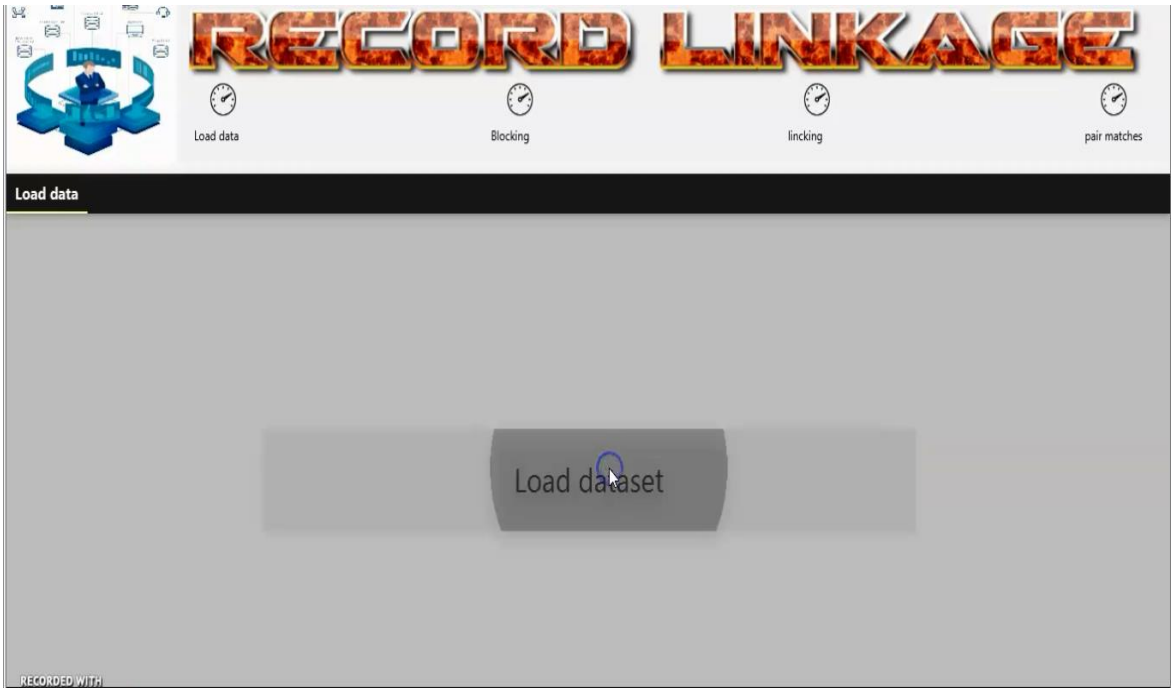


FIGURE 4.1 : Interface principale de notre application

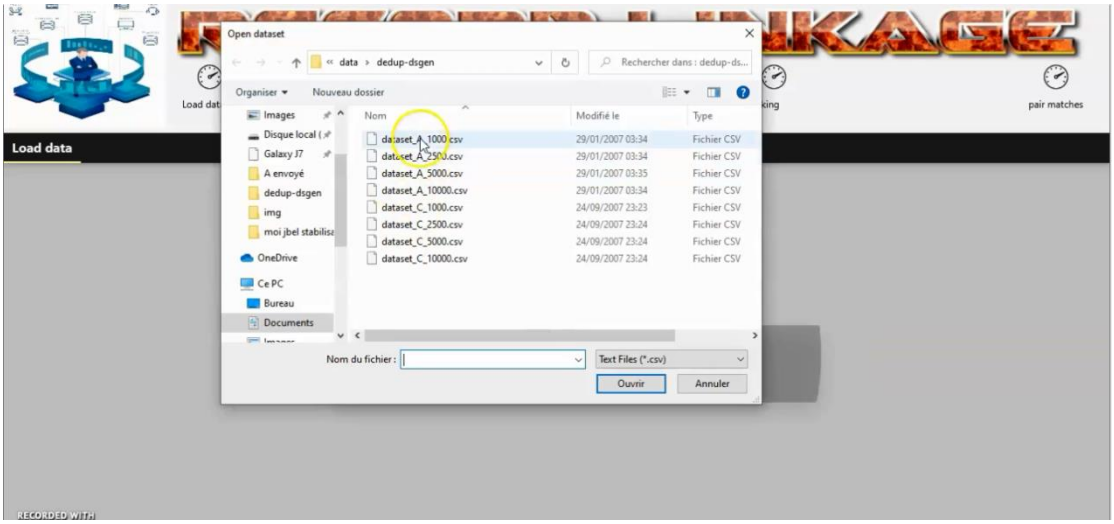


FIGURE 4.2 : Sélectionnes fichier data set

2- ensuite tous les fields est identifié par le fichier est catte fields composè par des block (FIGURE 4.3)

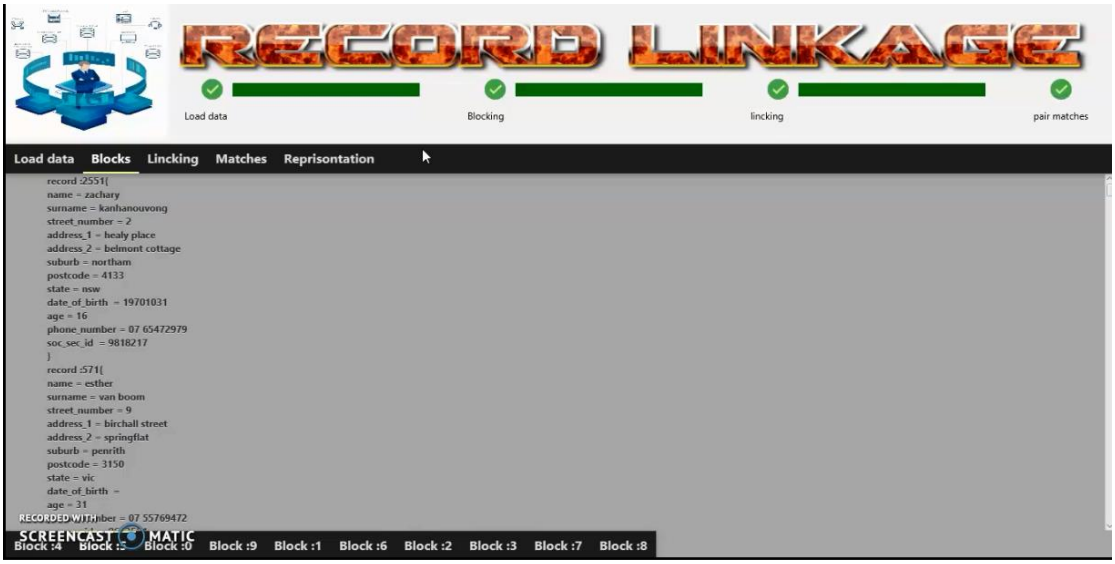


FIGURE 4.3 : identifier les fichier fields et La Classification des block

3- Dans cette étape, nous avons travaillé sur la création d'un enregistrement (pair record) car le nombre de dataset est grand et l'ordinateur ne peut pas le calculer comme ce que nous avons mentionné dans le chapitre précédent, et à partir de là, nous l'avons divisé en pages qui sont meilleures que de travailler sur une page pour résoudre le problème de calcul(FIGURE 4.4)

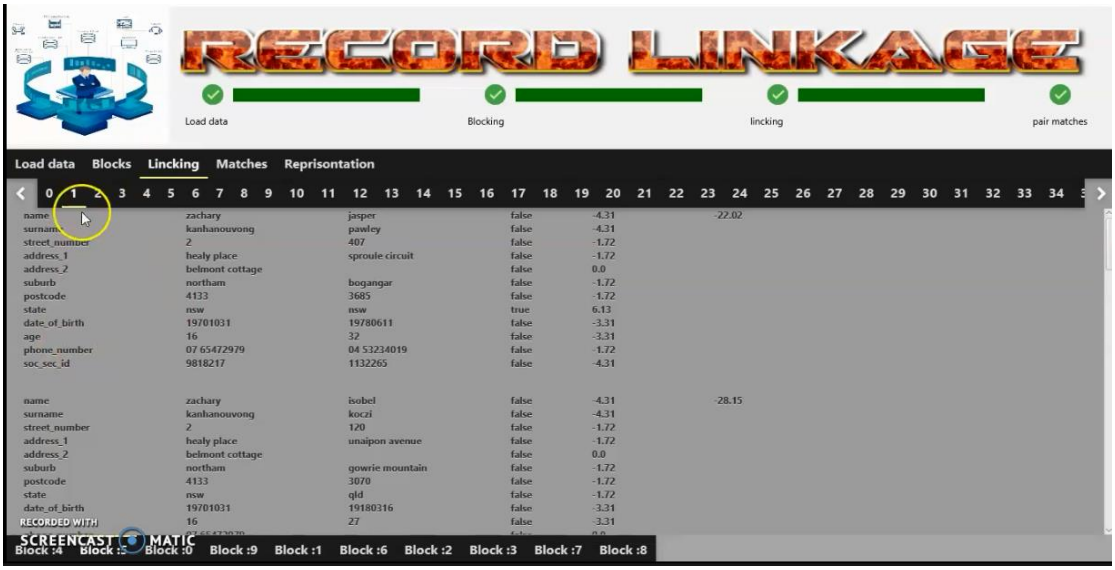


FIGURE 4.4 : la création d'un enregistrement (pair record)

4- tous les wieghtseuil Pour évaluer le traitement et la performance de similarité de deux enregistrement (2 Record) afficher out ça dans la fenêtre " match "

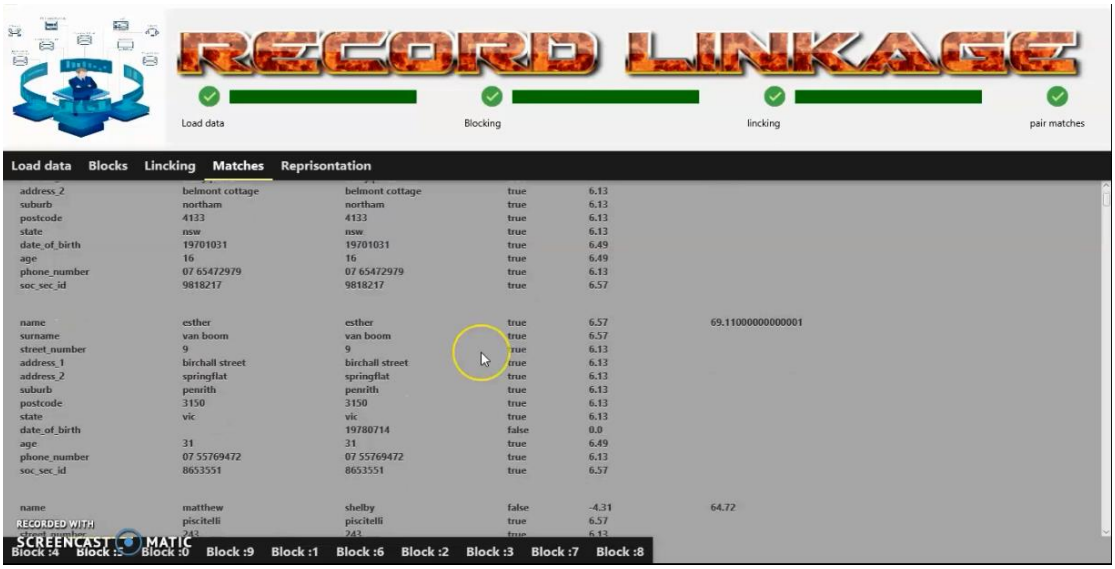


FIGURE 4.5 : affichage de résultat deux record sur le match

5- L'affichage de résultat de deux record nous avons avoir la classification

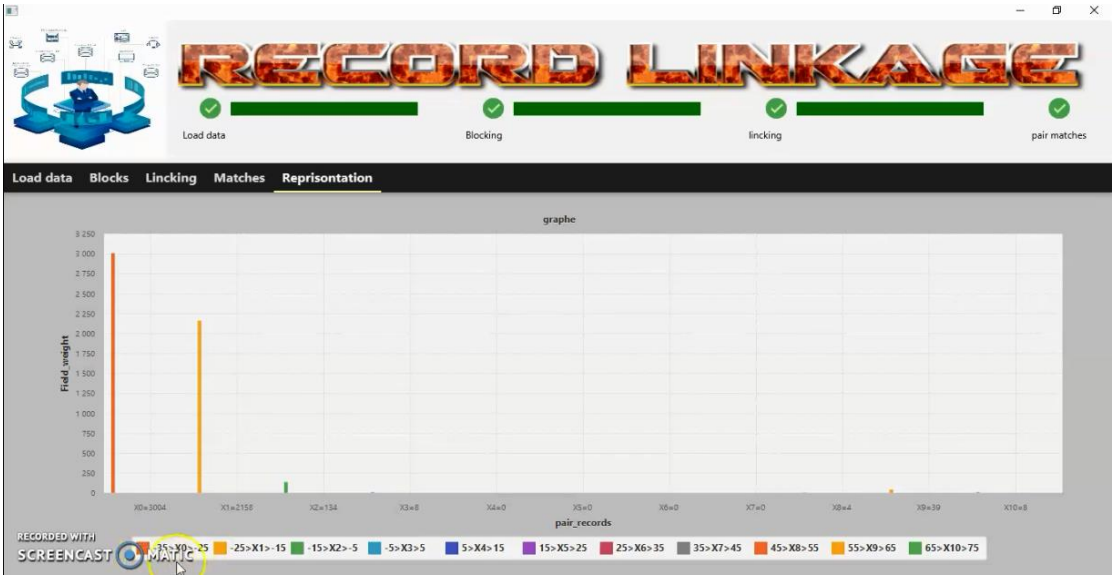


FIGURE 4.6 : Affiche les résultats du record

4.5 Conclusion

Ce chapitre détaille le processus de couplage d'enregistrement probabiliste l'approche basé sur les règles utilisé. Nous avons consacré à la réalisation et implémentation des différentes procédures, ainsi que l'évaluation du résultat obtenu.

CONCLUSION GÉNÉRALE

Résolution d’entités est définie comme l'identification de différentes descriptions ou enregistrements qui représentent la même entité du monde réel. Ces représentations sont considérées comme des doublons ou similaires dues à des erreurs et à des incohérences dans les données, telles que des fautes de saisie et des fautes d'orthographe, des informations manquantes ou des données obsolètes. La résolution d’entité est une étape très importante dans les processus de nettoyage et d'intégration des données qui permet d’améliorer la qualité des données.

Dans les cas où un identifiant unique est manquant, le couplage d'enregistrements s'est avéré être un outil utile, en particulier dans les situations où le chevauchement entre les ensembles de données à fusionner est important et la quantité de bruit dans les données est modérer.

Dans le Couplage d'enregistrements probabilistes, chaque variable de liaison a un certain poids. Le poids global des variables de couplage est utilisé pour décider si une paire d'enregistrements correspondante peut être liée ou non,

Notre implémentation montre que les résultats obtenus par le couplage d'enregistrements probabilistes sont satisfaisants et encourageants.

BIBLIOGRAPHIE

[1] Qualité contextuelle des données : détection et nettoyage guidés par la sémantique des données

[2] Dspace Qualité de données pour l'intégration de données THESE DE DOCTORAT EN SCIENCE **OUHAB Abdelkrim** sidi bel abbès 2018-2019

[3] Dspace Belghoul Badrddine, Qualité des donnes dans un Data warehouse,Elmagarmid,Memoire de Master UNIVERSITE KASDI MERBAH OUARGLA,algerie (2014/2015).

[4] Louardi Bradji, Adaptation des techniques de l'Extraction des Connaissances partir des Donnes (ECD) pour prendre en charge la qualit des donnes,Thse de Doctorat en Informatique. Universit Mentouri Constantine, (Mars 2012).

[5] Dibben, C., Williamson, L., Huang, Z.: Digitising Scotland (2012), http://gtr. rcuk.ac.uk/projects?ref=ES/K00574X/2

[6] Using Metric Space Indexing for Complete and Effcient Record LinkageSchool of Computer Science, University of St Andrews, St Andrews,Scotland.Contact:fozgur.akgun, alan.dearle, graham.kirbyg@st-andrews.ac.ukResearch School of Computer Science, The Australian NationalUniversity.Canberra,Australia.Contact:peter.christen@anu.edu.au

[7] ihel Kooli. Rapprochement de données pour la reconnaissance d'entités dans les documents océrisés. Intelligence artificielle [cs.AI]. Université de lorraine, 2016. Français. tel-01515422v1

[9] Aïcha Ben Salem. Qualité contextuelle des données : détection et nettoyage guidés par la sémantique des données. Performance et fiabilité [cs.PF]. Université Sorbonne Paris Cité, 2015. Français. NNT :2015USPCD054. tel-016624

[10] Recueil du Symposium 2016 de Statistique Canada Croissance de l'information statistique : défis et bénéfices (Aperçu du couplage d'enregistrements de données d'entreprises à Statistique Canada : Comment coupler les enregistrements « non couplables ») Javier Oyarzun et Laura Wile1

[11] https://thodrek.github.io/CS839_spring18/papers/Fellegi69.pdf

[12] Memobust Handbook on Methodology of Modern Business Statistics Probabilistic Record Linkage [26 March 2014]

[13] Data Matching Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection [2012]

[14] Data Quality and Record Linkage TechniquesWilliam E. Winkler, Thomas Herzog, Fritz J. Scheuren

[15] Record Linkage Methods with Applications to Causal Inference and Election Voting Data by Joan Pearson Heck Wortman Department of Statistical Science Duke University Dissertation submitted in partial ful_llment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Sciencein the Graduate School of Duke University 2019

[16] Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia .., Partie 3

[17] IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. Z, NO. Y, ZZZZ 2011 1 A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication

[18]https://dspace.univ-ouargla.dz/jspui/bitstream/123456789/1617/1/Master-Ahfouda-Habbi.pdf

[19]https://www.labri.fr/perso/johnen/pdf/IUTBordeaux/UMLCours/IntroductionJavaFX-V1.pdf