الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التعليم العالي والبحث العلمي

جامعة سعيدة د. مولاي الطاهر

كلية التكنولوجيا

قسم: الإعلام الآلي

Mémoire de Master

Spécialité : Réseaux Informatique et Systèmes Réparties

Thème

La sélection des attributs lors de la mise en correspondance (matching) pour le couplage d'enregistrements

Présenté par :

Sayah Mohammed Mouaad

Louazani Chahra zed

Dirigé par :

Benyahia Miloud

Promotion 2021 - 2022

Remerciements

Au nom d'ALLAH
Le Clément et le Miséricordieux
On trouve dans la tradition prophétique le hadith : «Celui qui ne remercie pas lesgens n'a pas remercié Allah »

Je tiens tout d'abord à exprimer ma sincère gratitude au Monsieur Benyahia Miloud, pour la confiance qu'il a bien voulu m'accorder en acceptant de dirigerce mémoire pour la qualité de son encadrement, ses précieuses orientations, sa simplicité .

Mes plus sincères remerciements : Aux membres du jury ; qu'ils soient remerciésde nous avoir fait l'honneur de juger notre travail.
Je remercie tous ceux qui m'ont aidée de près ou de loin à l'élaboration de travailde recherche.

Dédicace :

J'ai le plaisir de dédier ce travail :

A ma chère mère qui me donne toujours espoir pour la vie et qui ne cesse de prier pour moi.

A mon très cher père pour ses encouragements et son soutien, surtout pour son amour et son sacrifice jusqu'à rien

Cela n'interfère pas avec le déroulement de mes études

A mes amis les plus chers : Sadi Ossama et mes chers collègues

Tout ce qui m'aide et m'oblige à faire ce travail

Enfin, je tiens à remercier mon associé LOUAZANI qui a contribué à la réalisation de ce travail

SAYAH MOHAMMED MOUAAD

Dédicace :

J'ai le plaisir de dédier ce travail :

A ma chère mère qui me donne toujours espoir pour la vie et qui ne cesse de prier pour moi.

A mon très cher père pour ses encouragements et son soutien, surtout pour son amour et son sacrifice jusqu'à rien

Cela n'interfère pas avec le déroulement de mes études

A mes amis les plus chers : Karima et Naceur mes chers collègues

Tout ce qui m'aide et m'oblige à faire ce travail

Enfin, je tiens à remercier mon associé SAYAH qui a contribué à la réalisation de ce travail

CHAHRAZED LOUAZANI

الخلاصة

تعد عملية تحديد أزواج السجلات التي تمثل نفس الكيان الواقعي في قواعد بيانات متعددة ، والتي يشار إليها عادةً باسم ارتباط السجل أو ارتباط السجل (RL) ، إحدى الخطوات الأولية المهمة في العديد من تطبيقات معالجة البيانات.تعدين البيانات.

يمكن تعريف ربط السجل بأنه عملية من ثلاث خطوات: (1) التنظيف والتطبيع (2) الفهرسة والحظر (3) مطابقة أزواج السجلات المفهرسة (المطابقة).

كسمة واحدة حيث تتحكم معلمتان (BK: Blocking Key) يمكن اختيار مفتاح الحظر (BKV: مهمتان في أداء مفاتيح الحظر بتسلسل العديد من السمات: قيمة مفتاح الحظر و عدد مفاتيح الحظر ويتم استخدامه خلال المرحلتين الأخيرتين (قيمة مفتاح الحظر من RL.

تشمل خصائص السمات التي تؤثر على قرار الاختيار مستوى الأخطاء في قيم السمات وعدد (وتوزيع) قيم السمة ، أي محتوى معلومات السمة.

أظهرت النتائج التي تم الحصول عليها من التجارب على مجموعات البيانات الواقعية كفاءة اختيار مفاتيح الحظر المختلفة في كل خطوة من خطوات RL.

Résumé

Le processus d'identification des paires d'enregistrements qui représentent la même entité du monde réel dans plusieurs bases de données, communément appelé couplage d'enregistrements ou le record linkage (RL), est l'une des étapes initiales importantes de nombreuses applications d'exploration de données.

Le record linkage peut être défini comme un processus en trois étapes : (i) Le nettoyage et la normalisation (ii) L'indexation et le blocage (iii) La mise en correspondance des paires d'enregistrements indexés (Matching).

Une clé de blocage (BK: Blocking Key) peut être choisie comme un attribut unique où avec la concaténation de plusieurs attributs, deux paramètres importants contrôlent les performances des clés de blocage : la valeur de la clé de blocage (BKV: blocking key value) et le nombre de clés de blocage et elle est utilisée durant les deux dernier étapes de RL.

Les caractéristiques d'attribut qui affectent la décision de sélection comprennent le niveau d'erreurs dans les valeurs d'attribut et le nombre (et la distribution) des valeurs d'attribut, c'est-à-dire le contenu informationnel de l'attribut.

Les résultats obtenus à partir des expériences sur des data sets du monde réel ont montré l'efficacité de choisir des clés de blocage différentes dans chaque étapes de RL.

MOTS CLÉS

Record linkage, clés de blocage, blocage, Matching, sélection des attributs.

ABSTRACT

The process of identifying pairs of records that represent the same real-world entity in multiple databases, commonly referred to as record linkage or record linkage (RL), is one of the important initial steps in many data processing applications. data mining.

Record linkage can be defined as a three-step process: (i) Cleaning and normalizing (ii) Indexing and blocking (iii) Matching pairs of indexed records (Matching).

A blocking key (BK: Blocking Key) can be chosen as a single attribute where with the concatenation of several attributes, two important parameters control the performance of the blocking keys: the value of the blocking key (BKV: blocking key value) and the number of blocking keys and it is used during the last two stages of RL.

Attribute characteristics that affect the selection decision include the level of errors in attribute values and the number (and distribution) of attribute values, i.e. the information content of the attribute.

The results obtained from the experiments on real-world datasets showed the efficiency of choosing different blocking keys in each RL step.

KEY WORDS

Record linkage, blocking keys, blocking, Matching, attribute selection.

INTRODUCTION GÉNÉRALE

- La plupart des systèmes d'information sont effectués par l'existence des doublons dans leur base de données, qui sont causé par différents facteur tel que l'intégration de différentes sources de donnée hétérogène et distribués. Cette anomalie provoquera des problèmes tels que la dégradation des performances, l'augmentation des couts et la manque de qualité. Cela être éviter par la procédure de déduplication des enregistrements se réfère à l'identification de la même entité du monde réel avec des différentes représentations.

  L'objective de notre travail : nous allons intéresser au problème de qualitéde données des données pour les données en doubles et la façon de résoudre ce problème par un processus qui vient d'être décrit a été mis en application afin d'évaluer et comparer les performances en utilisant plusieurs et différents clé de blocage dans l'étape La mise en correspondance des paires d'enregistrements indexés (Matching).

Le mémoire est structuré en IV 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 allons présenter le Recorde Linkage, L'indexation et le blocage et la mise en correspondance des paires d'enregistrements indexés (Matching).

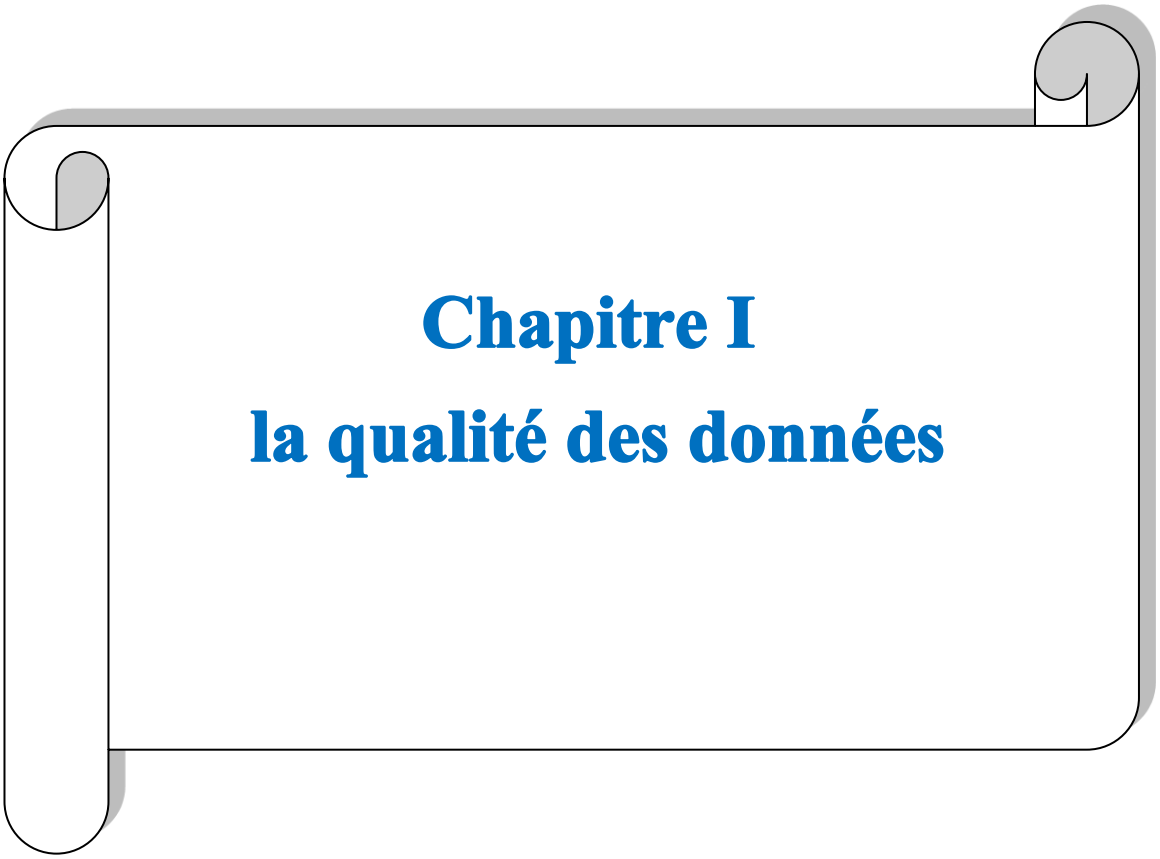
Dans le troisième chapitre :

Nous présentons la conception et la modélisation de notre système et architecture pour la sélection des attributs et la construction des clés de blocage (BK) qui seront utilisées dans l'étape d'indexation (blocage) et matching.

Dans le quatrième chapitre :

Nous présentons l'implémentation évaluons Algorithme de K-Mods sur des bases de données réelles et nous terminons par les tests effectues et les résultats obtenus.

Table des matières



Chapitre I
la qualité des données

I - Introduction

De nos jours, avec les développements technologiques, les entreprises stockent de plus en plus de donnée. Malheureusement les travaux de maintenance et de la qualité des données sont souvent négligés, pourtant les données de mauvaise qualité constipée tuent un facteur de cout important.

Les données de mauvaise qualité peuvent donc avoir des effets significativement négatifs sur l'efficacité d'une organisation, alors que les données de qualité, sans doublons, sans anomalies sont souvent essentielles au succès d'une entreprise.[1]

Dans ce chapitre, nous commencerons par définir qualité des données et leur concept, par la suite on abordera les conséquences de la non-qualité. Ensuite on citera les dimensions et les critères de la qualité des données et aussi citera l'amélioration de la qualité de donnée et la duplication de donnée. Et finalement on parlera sur l'objectif de la qualité de donnée.

I.1 La qualité de données :

I.1.1 Définition 1 de la qualité:

-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 dès le début des normes et standards internationaux, nous nous intéressons ici aux définitions don- nées par l'organisation internationale de standardisation (ISO : International Standard Organisation) et par Organisation de Coopération et de Développement Economiques (OECD : Organisation for Economique Coopération and Développent).[1]

a.La Définition de ISO : La norme ISO définit la qualité comme "L'ensemble des propriétés et caractéristiques d'un produit ou d'un service qui lui confère l'aptitude à satisfaire des besoins exprimés ou implicites". Pratiquement, la qualité d'un produit signifie qu'il est adapté au besoin qu'il est censé satisfaire. La notion de qualité s'applique aussi bien à des produits qu'à des services.

b.La Définition de 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.[2]

I.2 La définition 2 de la qualité de données:

La qualité des données est un terme générique décrivant à la fois les caractéristiques des données : complète, fiable, pertinents, cohérente set à jour, il permet de garantir ces caractéristiques par des ensembles des processus.[3]

La qualité des données est pour le but d'obtenir les données sans doublons, sans fautes d'orthographe, sans omission, sans variation superflue et conforme à la structure, il fait aussi référence à l'utilité globale d'un ou de plusieurs jeux de données en fonction de sa capacité à être facilement traitée et analysée pour d'autres utilisations, généralement par une base de données, un entrepôt de données ou un système d'analyse de données.

I. 3 Les critères de la qualité des données :

I.3.1 Les critères intrinsèques :

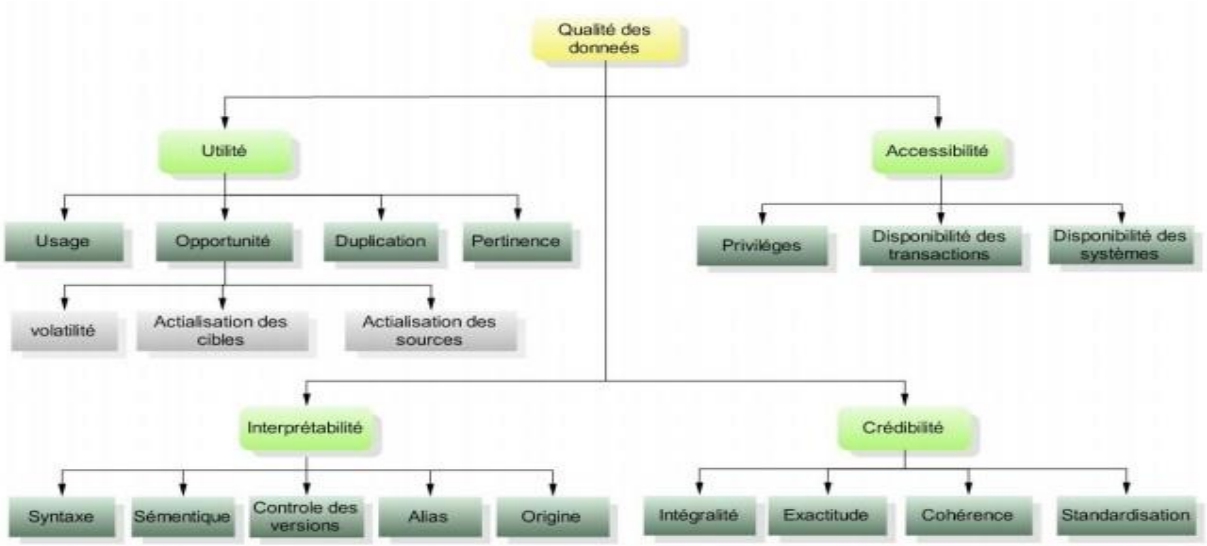


FIGURE 1.1: Les dimensions de la qualité des données

a- L'unicité:

L'unicité des données est un aspect de la qualité des données qui désigne le résultat des processus visant à résoudre et à éviter les problèmes de duplication indésirable des données.

L'unicité des données sert aussi à n'avoir qu'une seule description d'un produit donné.

Elle contribue alors à l'amélioration de la qualité des données produit.[4],

b- L'exactitude:

Une donnée est " exacte " si la valeur des attributs de l'entité concernée est égale à la grandeur qu'elle est censée représenter dans le monde réel. Cette notion englobe donc deux aspects : la précision et la validité.[4]

c- La complétude:

La complétude des données possédez-vous toutes les données dont vous avez besoin sur vos clients ou prospects ? Tous les attributs / champs dont vous avez besoin sont-ils renseignés ? L'incomplétude des données à disposition est l'un des principaux challenges rencontrés par les organisations dans leur gestion des bases de données.

d- La cohérence :

Cette notion est relative à l'absence d'informations conflictuelles au sein d'un même objet (par exemple, une incohérence serait détectée si un " prix actuel " d'un produit est supérieur au " prix maximum " de ce même produit). Mais cette notion existe aussi au niveau service : les valeurs d'une instance d'un objet métier ne sont pas en conflit avec les valeurs d'une autre instance ou d'une instance d'un autre objet.[5],

e- L'intégrité:

Toutes les données nécessaires sont disponibles pour le besoin de métier. Il est impossible d'effectuer une compagne d'e-mailing avec une base de données clients ne contenant pas l'adresse mail.

I. 4 Les Critères de Sécurité à L'actualité :

Quand est-ce que les données ont été enregistrées ou mises à jour ? On en revient à ce que nous disions tout à l'heure : les données perdent naturellement de la valeur avec le temps, elles se dégradent. Plus une donnée est récente, toutes choses égales par ailleurs, plus elle a de valeur.

a-L 'accessibilité :

Est la dimension qualité qui concerne la facilité d'accès aux données. Cela signifie que les services de données sont calibrés en fonction de leur utilisation et qu'ils existent souvent aussi bien en mode événement (déclenché à chaque mise à jour), qu'en mode requête (à la demande d'un processus consommateur) ou en mode batch pour des synchronisations en masse (pour le décisionnel par exemple).[5]

b- La pertinence:

La pertinence est la dimension qualité qui définit l'utilité d'une donnée. Une donnée peut être accessible mais tellement détaillée que de nombreux attributs de l'objet pro- posé sont inutiles aux processus consommateurs. Une donnée doit être adéquate à son usage. Les services de donnée seront d'autant mieux utilisés que la granularité d'information dispensée correspondra aux besoins.[5]

c- La compréhensibilité :

La compréhensibilité est la dimension qualité associée à la question : " cette don- née est-elle compréhensible ? ". Une donnée est compréhensible si chaque utilisateur, chaque processus, chaque application trouve facilement la bonne information parmi les attributs disponibles d'un objet. C'est le cas si celui-ci est clair et que l'aligne- ment sémantique de l'ensemble des concepts entre tous les dépositaires (humains ou informatiques) a été réalisé et documenté.[5]

I.5 Principaux problèmes du non qualité des données

I.5.1 Les problèmes de la qualité des données :

Pour définir les problèmes de qualité dans l'entreprise, il est recommandé de définir les dimensions possibles et leur importance :[5]

a-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.

b-Standards : les valeurs sont correctes par rapport à un intervalle de répartition ou à un domaine.

c-Par manque de standards de codification : l'entreprise " Les chantiers Techniques de Marseille" peut apparaître comme $< c. EtsCTM > < cCTM >$, où $< CTMSA >$.

d-Intégralité : toutes les données nécessaires sont disponibles pour le besoin métier. Il est impossible d'effectuer une campagne d'e-mailing avec une base de données clients ne contenant pas l'adresse email.

e-Exactitude : les données représentent la réalité ou sont vérifiables à partir d'une source externe - Le code postal ne correspond pas à la localité, le téléphone a changé ou le SIRET n'a pas été mis à jour lors du déménagement de l'entreprise.

f-Interopérabilité : une donnée doit être représentée sous un format cohérent et sans ambiguïté.

g-Opportunité : les données sont à jour au moment de leur utilisation. Le rapport mensuel des ventes doit inclure tous les résultats actualisés du mois pour toutes les régions commerciales.

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.

I. 5.2 Problème de rencontres:

Les problémes des données ne naissent pas de nulle part, les causes de la non-qualité des données sont connues : On trouve les problèmes techniques ou les problèmes humains.Ces probléme s'accumulent avec le temps, depuis la création, durantla manipulation et jusqu'à l'exploitation et l'analyse.[6]

A. Création des données :

la création des données passent par la conception et la modélisaton de la base de données jusqu'à l'entrée des données par les utilisateurs, durant cet étape de création beaucoup de causes ou sources de problémes ont été relevées :

1) Les Entrée manuelle : absence de vérifications systématiques des formulairesde saisie.
2) Entrée automatique : problèmes de capture OCR, de reconnaissance de la pa-role Incomplétude, absence de normalisation ou inadéquation de la modélisation conceptuelle des données attributs peu structurés, absence de contraintes d'intégrité pour maintenir la cohérence des données.
3) Entrée de doublons.
4) Approximations.
5) Contraintes matérielles ou logicielles.
6) Erreurs de mesure.
7) Corruption des données : faille de sécurité physique et logique des données.

B. Collecte Import des données

1) Destruction ou mutilation d'information par des prétraitements inappropriés.
2) Perte de données : buffer overflows, problèmes de transmission.
3) Absence de vérification dans les procédures d'import massif.
4) Introduction d'erreurs par les programmes de conversion de données.

C. Stockage des données

1) Absence de méta-données.
2) Absence de mise à jour et de rafraîchissement des données obsolètes ou répliquées.
3) Modifications ad-hoc.
4) Modèles et structures de données inappropriés, spécifications incomplètes ou évolution des besoins dans l'analyse et conception du système.
5) Contraintes matérielles ou logicielles.

D. Intégration des données

1) Problèmes d'intégration de multiples sources de données ayant des niveaux de qualité et d'agrégation divers.
2) Problèmes de synchronisation temporelle.
3) Systèmes de données non conventionnels.
4) Facteurs sociologiques conduisant à des problèmes d'interprétations et d'intégration des données.

E. Recherche et analyse des données

1) Erreur humaine.
2) Contraintes liées à la complexité de calcul.
3) Contraintes logicielles, incompatibilité.
4) Problèmes de passage à l'échelle, de performances et de confiance dans les résultats.
5) Approximations dues aux techniques de réduction des grandes dimensions.

6 Amélioration de la qualité des données

6.1.Approche globale

L'amélioration de la qualité des données est une démarche continue. Elle commence dès l'analyse des sources de données, et se poursuit avec la préparation du chargement du référentiel, et consiste enfin en un suivi régulier de l'activité.[7]

a. Les étapes d'améliorer la qualité des données :
b. Valider le niveau de qualité sur l'existant.
c. Définir le niveau de qualité cible.
d. Atteindre le niveau de qualité cible.
e. Rester à ce niveau.
f. Surveiller la qualité.

6.2 Approche " nettoyage ", ou data cleansing :

Le nettoyage de données est l'opération de détection et de correction (ou suppression) d'erreurs présentes sur des données stockées dans des bases de données ou dans des fichiers.

Le problème de nettoyage des données qui consiste à détecter et éventuellement corriger des incohérences et des erreurs trouvées dans des jeux des données originaux, est bien connu dans le domaine de l'aide à la décision et des bases des données . Le nettoyage des données est un processus itératif et interactif qui comporte trois phases.[7]

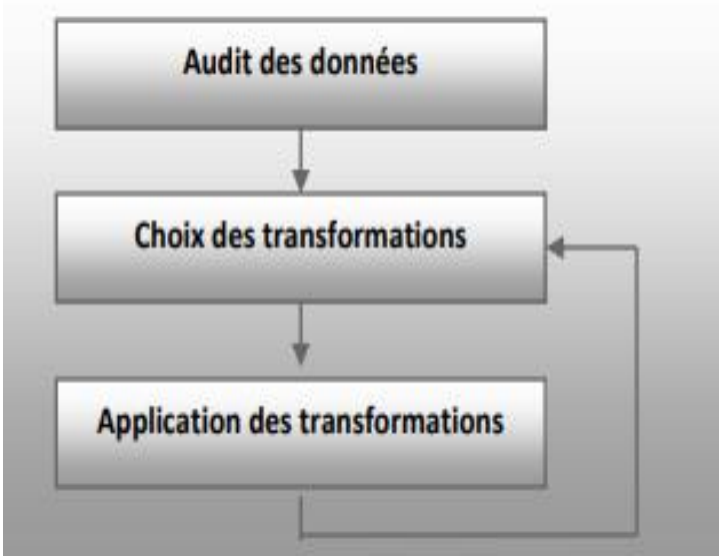


FIGURE 1.2: Processus du nettoyage des données

6.3 Approche " processus":

L'approche " processus " a pour objectif de prévenir l'introduction de données erronées dans un système d'information. On entend par " processus " toute la chaîne de traitements et d'opérations, de la création des données à leur destruction, en passant éventuellement par des modifications de leurs valeurs.[7]

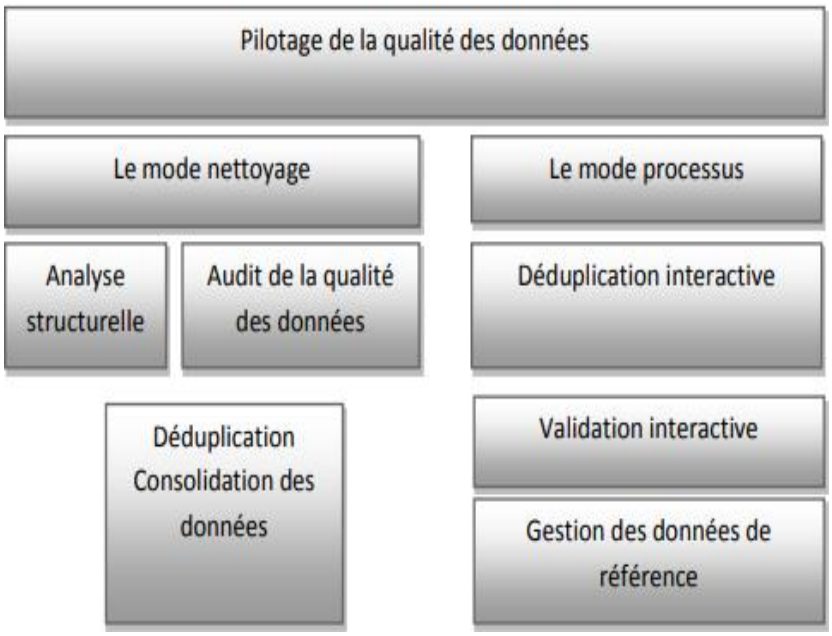


FIGURE 1.3: Approches " nettoyage " et " processus " de la qualité des données

7. Les objectives de la qualité des données:

1- La qualité des données se rapporte à des informations exactes et fiables collectées par l'intermédiaire d'un système de gestion des données de suivi et d'évaluation(S,E).[6]

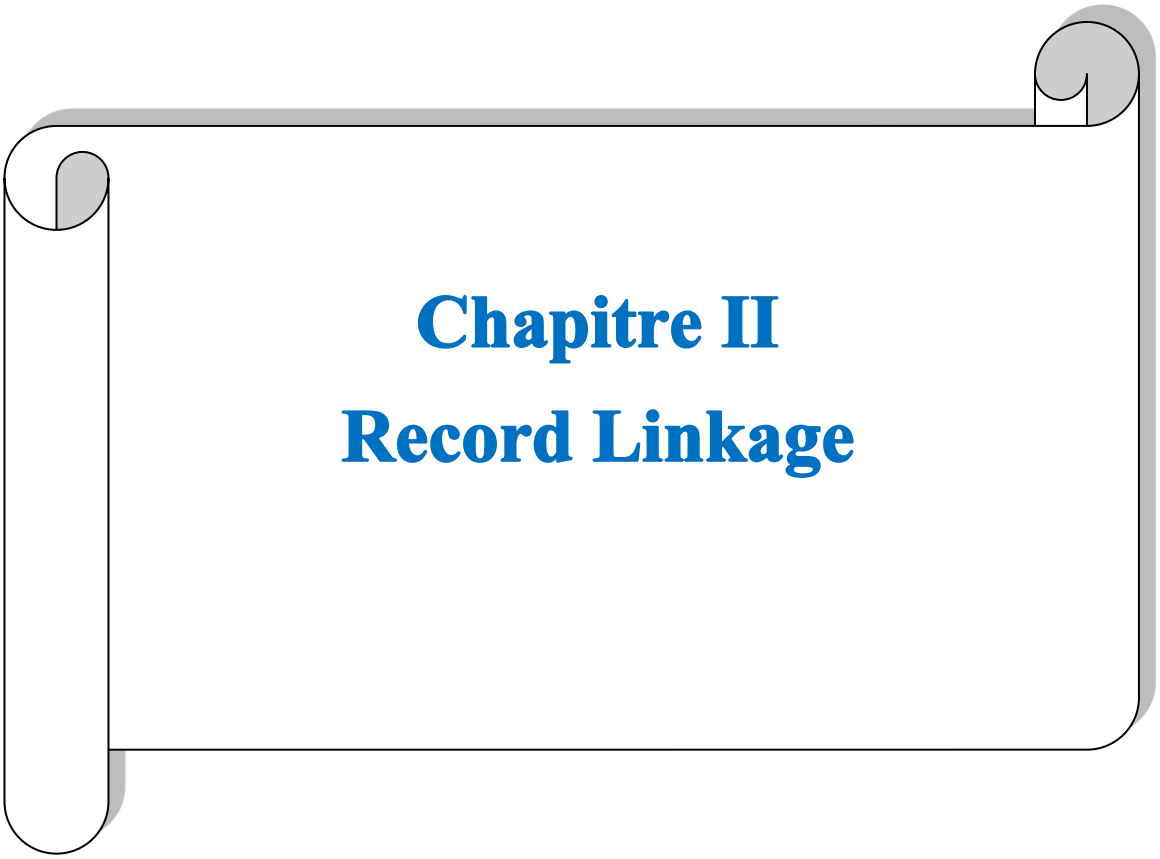
2- La qualité des données est importante pour les programmes sur le VIH-sida car ils sont généralement orientés sur les résultats.

3- Les données de qualité sont importantes pour surveiller et évaluer les progrès réalisés pour atteindre ces objectifs.[6]

8 Conclusion:

Dans ce chapitre, nous avons présenté un état de l'art sur la qualité des données et après, on situe les critères de la qualité et les différents problèmes de la non-qualité en après on posé comment améliorer la qualité des données.

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



Chapitre II
Record Linkage

II. 1 Définition

Record Linkage (RL), est le processus d'identification des tuples qui font référence à la même entité du monde réel. Sans blocage, le processus RL peut aboutir à des milliards de comparaisons lorsqu'il s'agit de grands ensembles de données [8]

Il existe deux méthodes principales de record linkage [8]:

A- Déterministe : telle que celle proposée dans [Lee00], se basent sur des règles définies par des experts déterminant les conditions de couplage d'une paire d'enregistrements. Ces règles sont gé-néralement 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. 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.[9]

D'autres techniques [Church91, Bitton83, Kukich92] ramènent la comparaison à des mesures de si milarité entre les attributs pour corriger les variations de représentation, comme la mesure de Levenshtein [Monge97]. En revanche, ces techniques obligent également l'utilisateur à intervenir pour fixer les seuils de mesures.[10]

B- probabiliste : il 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. Elles peuvent être non supervisées ou supervisées.[11]

II 2 Les étapes de Record Linkage

Le record linkage peut être défini comme un processus en trois étapes :

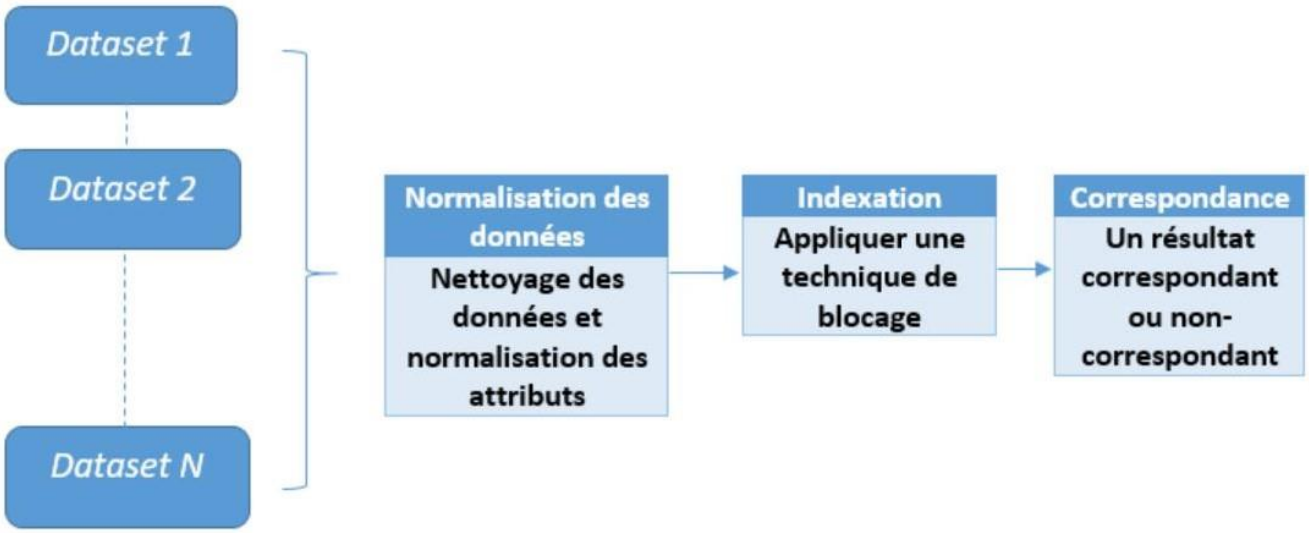


Figure 2.1 – Les étapes de Record Linkage.

II.2.1 . Nettoyage et normalisation : L'application du processus RL à des données corrompues peut entraîner la fusion de tuples incorrects et la perte d'informations importantes dans la base de données. Par exemple, l'attribut d'adresse peut être représenté comme un seul champ dans une base de données et comme plusieurs champs dans une autre base de données (code postal, rue, ville, etc.). Par conséquent, pour faciliter le processus RL, vous devez effectuer une normalisation du champ d'adresse avant de démarrer le processus RL.[12]

II.2.2 L'indexation : Ceci est considéré comme l'étape la plus importante de ce processus. L'indexation combine tous les enregistrements de correspondance possibles dans le même bloc à des fins de comparaison. La technique d'indexation la plus courante est le "blocage".[13]

II.2.2.1. K- Modes

II.2.2.1.1 Présentation de la méthode:

Les K-Modes ont été proposés pour la première fois par HUANG en 1998 [15]. Cet algorithme est considéré comme une extension de l'algorithme de cluster traditionnel -K-Means, et il est proposé de regrouper les données de catégorie , contrairement à l'algorithme -K-Means, qui n'accepte que l'attribut des nombres. Bien sûr, il existe des algorithmes de clustering hiérarchiques traditionnels qui gèrent à la fois les données catégorielles et numériques, mais leur complexité secondaire les rend inadaptés au clustering de grands ensembles de données. En utilisant le mode K dans notre approche, nous avons pu éliminer l'étape de conversion des données numériques . C'était une étape très longue et nécessaire pour l'algorithme des k-moyennes. L'auteur de a créé l'algorithme basé sur trois

$$d_1(X,Y) = \sum_{j=1}^{m} \delta(x_i, y_i), \quad \alpha = \begin{cases} 0 & (x_i = y_i) \\ 1 & (x_i \neq y_i) \end{cases} \qquad (1)$$

points principaux [14]:

(1) Une simple mesure de dissemblance de qui correspond à un objet

(2) (prouvé par l'équation 1). (2) Utilisation modes à la place des moyens et (3) Approche basée sur la fréquence pour mesurer le mode d'un ensemble.

Deux méthodes ont été proposées dans le magazine pour la sélection initiale des modes.[14]

La première consiste à attribuer le premier enregistrement individuel K comme mode initial. La deuxième approche consiste à mesurer les comptes pour les catégories pour chaque attribut et à les trier par ordre décroissant de comptes. Une fois cela fait, attribuez

la première catégorie de fréquence au premier mode k initial. Dans notre travail, nous utilisons **l'algorithme K-Modes** pour regrouper les données en blocs. Chaque bloc contient correspondances possibles. Au lieu d'utiliser tous les attributs du jeu de données,

le cluster utilise uniquement la clé de blocage générée à l'étape précédente comme attributs du cluster[14]

II.2.2.2 Le blocage:

II.2.2.2.1 Définition:

Le blocage est la technique la plus utilisée dans l'étape del'indexation.

Le blocage est le processus qui divise le dataset en unensemble de blocs.[15]

Tous les tuples affectés au même bloc partagent une valeur commune appelée valeur de clé de blocage (BKV). La clé de verrouillage peut être sélectionnée comme un attribut unique.[16]

Par exemple, tous les enregistrements qui partagent la même valeur pour l'attribut d'adresse sont affectés au même bloc. Vous pouvez également concaténer plusieurs attributs pour sélectionner la clé de blocage, tels que les quatre premières lettres du nom ou le code postal de l'attribut d'adresse.

BK	Name	Address	City	Phone	TYpe
Losangelos310/246-1501	Amine morton's	435 s.la ceine ga blv	Los angelos	310/246-1501	Americ an
Studiocity818/762-1221	Art's delicatess en	12224 ventura blvd	Stud io city	818/762-1221	Americ an

Tableau II.2 – Exemple de clé de blocage

II.2.2.2.2 Codage phonétique :

a- Soundex :

Les principales étapes du Soundex sont :

— Conservez la première lettre de la chaîne.

— Remplacez toutes les consonnes en utilisant les règles suivantes : (0 pour les caractères A, E, H, I, O, U, W, Y. 1 pour les caractères B, F, P, V. 2 pour C, G, J, K, Q, S, X, Z. 3 pour D, T. 4 pour L et 5 remplace M, N. 6 remplace le

caractère R.

— Dans le cas où la chaîne est trop courte, l'algorithme complète les trois chiffres après le premier caractère par des zéros.[17]

b-NYSIIS (Système d'identification et de renseignement de l'État de New York) :

Les règles de base de l'algorithme NYSIIS sont la transformation des premiers caractères où : (MAC est remplacé par MCC et KN devient NN, K en C, PH-PF en FF, SCH en SSS) et les derniers caractères (EE-IE en Y, DT-RT- RD-NT-ND en D). [17]

II.2.2.2.3 Recherche de motifs :

a-Distance d'édition :

elle est définie comme le nombre d'insertions, de suppressions et de mises à jour nécessaires pour transformer une chaîne en une autre. calculer les coûts de passage d'un mot à un autre. [17]

b-Jaro-Winkler

$$Jaro\_Sim(s_1, s_2) = \begin{cases} 0, & m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right), & otherwise \end{cases}$$

— **S :** représente la longueur de la chaîne.

— **m :** représente le nombre de caractères communs entre les séquences comparées avec le même indice.

— **t** : représente le demi nombre de transpositions.

Afin d'améliorer la métrique précédente, William E. Winkler utilise une échelle de préfixes P afin de privilégier les chaînes de caractères qui commencent par le même préfixe L pour une longueur maximale de quatre. La similarité de Jaro-Winkler est définie comme suit [17]:

Jaro Winlker_Sim (s_1 , s_2) = *Jaro_Sim*(s_1 , s_2) + *LP*(1 - *Jaro_Sim*(s_1 , s_2)) Où

— *Jaro_Sim* (s_1 , s_2) est la similarité Jaro entre les chaînes de caractères.

— L est la longueur du préfixe.

— P est un facteur d'échelle (une constante qui prend généralement la valeur (0,1).

c- distance de Jaccard:

La distance de Jaccard est généralement utilisée pour mesurer la similarité entre deux ensembles d'échantillons, Pour mesurer la distance de Jaccard, il faut d'abord calculer le coefficient de Jaccard qui est défini comme suit :

$$\textbf{Jaccard (A, B) = A} \cap \textbf{B /A} \cup \textbf{B}$$

Une fois cela fait, la distance de Jaccard est obtenue uniquement par la soustraction du coefficient de Jaccard de 1.[17]

$$\textit{Jaccard distance}\ (A, B) = 1 - \textit{Jaccard}\ (A, B)$$

II.2.2.3 La mise en correspondance des paires d'enregistrements indexés (Matching) : la dernière étape du processus de couplage d'enregistrements consiste à faire correspondre les enregistrements

indexés et à décider si deux paires comparées représentent ou non la même entité du monde réel. En général, la valeur de correspondance est normalisée dans la plage [0,1] où 1 représente une correspondance exacte et 0 une non-correspondance totale. Dans cette section, un bref aperçu des techniques de mise en correspondance existantes dans la littérature est fourni.

D'après la littérature, deux familles de techniques d'appariement existent. La première est le codage phonétique. L'idée de cette technique est de transformer une chaîne de caractères en un code qui représente la façon dont la chaîne est prononcée. Une variété d'algorithmes d'encodage phonétique existe dans la littérature (Soundex et phonex, phoenix, NYSIIS et Double-Metaphone). La deuxième famille de techniques d'appariement est la recherche de motifs. L'idée principale de ces techniques est de mesurer la similarité entre deux mots sans aucune transformation en utilisant un ensemble de mesures de similarité de chaînes de caractères telles que la distance d'édition. [17]

II. 2.2.3.1 Liens matching :

Les liens sont basés sur des caractéristiques similaires plutôt que sur

des informations d'identification uniques car des hypothèses fortes sur les relations communes sont faites . [18]

Les enregistrements liés ne doivent pas nécessairement correspondre à la même unité.

Étant donné que de nombreux travaux et le développement de logiciels associés ont été réalisés par différents groupes travaillant dans un isolement Dans l'espace produit A × B des fichiers A et B, une match est une paire qui représente la même entité commerciale et une non-match est une paire qui représente deux entités différentes. Avec une seule liste, un doublon

est un enregistrement qui représente la même entité commerciale qu'un autre enregistrement de la même liste. Plutôt que de considérer toutes les paires

de A × B, il peut être nécessaire de ne considérer que les paires qui concordent sur certains identifiants ou critères de blocage.

Une règle de décision de couplage d'enregistrements est une règle qui désigne une paire soit comme un lien, soit comme un lien possible, soit comme un non-lien. Les liens possibles sont les paires pour lesquelles les informations d'identification ne sont pas suffisantes pour déterminer si une paire est une correspondance ou une non-correspondance. En général, les commis examinent les liens possibles et décident de leur statut de matching

False matches sont soit non matches désignées comme non-liées par la règle de décision telle qu'elle est appliquée à un ensemble de paires,

soit des matches qui ne font pas partie de l'ensemble de paires auquel la règle de décision est appliquée. En général, lien/non-lien fait référence aux désignations en vertu des règles de décision et match/non-match fait référence au statut réel.[18]

Les principales raisons pour lesquelles les match sont utilisés pour l'appariement exact sontde réduire ou d'éliminer la révision manuelle et de rendre les résultats plus facilement reproductibles[18]

II.2.2.4 Conclusion :

Le processus d'identification des paires d'enregistrements qui représentent la même entité du monde réel dans plusieurs bases de données, communément appelé couplage d'enregistrements, est l'une des étapes initiales importantes de nombreuses applications d'exploration de données. Le couplage d'enregistrements de millions d'enregistrements est une tâche coûteuse en termes de calcul.

Chapitre III
analyse et La conception

III.1 Introduction :

Dans cette partie, on va analyser et modéliser les besoins du client avec le langage UML

L'activité d'analyse et de conception permet de traduire les besoins fonctionnels et les contraintes issues du cahier des charges et de la spécification des exigences dans un langage plus professionnel et compréhensible par tous les individus intervenants dans la réalisation et l'utilisation de l'application [10]

III.2 UML :

a. Choix D'UML:

UML, c'est l'acronyme anglais pour « Unified Modelin Language ». On le traduit par « Langage de modélisation unifié ». La notation UML est un langage visuel constitué d'un ensemble de schémas, appelés des diagrammes, qui donnent chacun une vision différente du projet à traiter. UML nous fournit donc des diagrammes pour représenter le logiciel à développer : son fonctionnement, sa mise en route, les actions susceptibles d'être effectuées par le logiciel, etc [10]

b. Pourquoi modéliser ?:

De la même façon qu'il vaut mieux dessiner une maison avant de la construire, il vaut mieux modéliser un système avant de le réaliser.

Modéliser, c'est décrire de manière visuelle et graphique les besoins, les solutions fonctionnelles et techniques du projet [10]

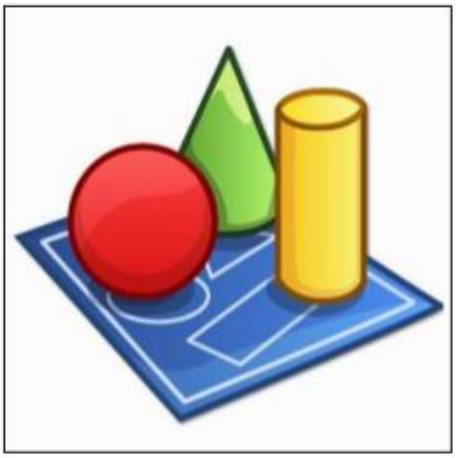
Modéliser pour :

.Obtenir une modélisation de très haut niveau indépendante des langages et des environnements.

Faire collaborer des participants de tous horizons autour d'un même document de synthèse.

Faire des simulations avant de construire un système.

III.3 Présentation des outils:



Power Designer (anciennement PowerAMC) est un logiciel de conception créé par la société SAP, qui permet de modéliser les traitements informatiques et leurs bases de données associées. [11]

PowerAMC propose différentes techniques de modélisation, chacune accessible aux informaticiens de tout niveau, parmi elles : Merise, UML, Data Warehouse, et processus métiers. Simple d'utilisation, personnalisable et dotée d'une interface intuitive, cette application optimise les productivités individuelle et collective. Elle intègre en outre des fonctions de génération de code pour plus de 45 bases de données et divers langages de programmation [11]

III.4 Identification des acteurs:

a) Acteur:

Un acteur est l'idéalisation d'un rôle joué par une personne externe, un processus ou une chose qui interagit avec un système.

Il se représente par un petit bonhomme (figure 1.) avec son nom (son rôle) inscrit dessous.

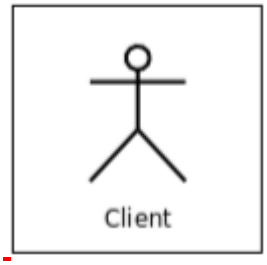


Figure 3.1 : Exemple de représentation d'un acteur

b) Cas d'utilisation :

Un cas d'utilisation est un service rendu à un acteur : c'est une fonctionnalité de son point de vue.

Un cas d'utilisation se représente par une ellipse (figure 2) contenant le nom du cas (un verbe à l'infinitif), et optionnellement, au-dessus du nom, un stéréotype. [12]

Figure 3.2 : Exemple de représentation d'un cas d'utilisation

c) Acteur direct :

Les acteurs directs, ce sont les utilisateurs de l'application, qui touchent directement l'application.

Type d'acteur	Description fonctionnelle
 Utilisateur	L'acteur le plus important, qui aura les fonctionnalités suivantes : - Téléchargez les Datasate et créez des clés Créez les bloks et Créez Matching - Etc…

Tableau III.1 : Acteurs primaires

III.5 Diagramme de cas d'utilisation:

Les diagrammes de cas d'utilisation sont des diagrammes UML utilisés pour donner une vision globale du comportement fonctionnel d'un système logiciel

Dans les figures qui suivent, nous présenterons les cas d'utilisation qui mettent en évidence les principales fonctionnalités de chaque acteur dans le système [12]

a . Diagramme de cas d'utilisation d'acteur «Utilisateur»:

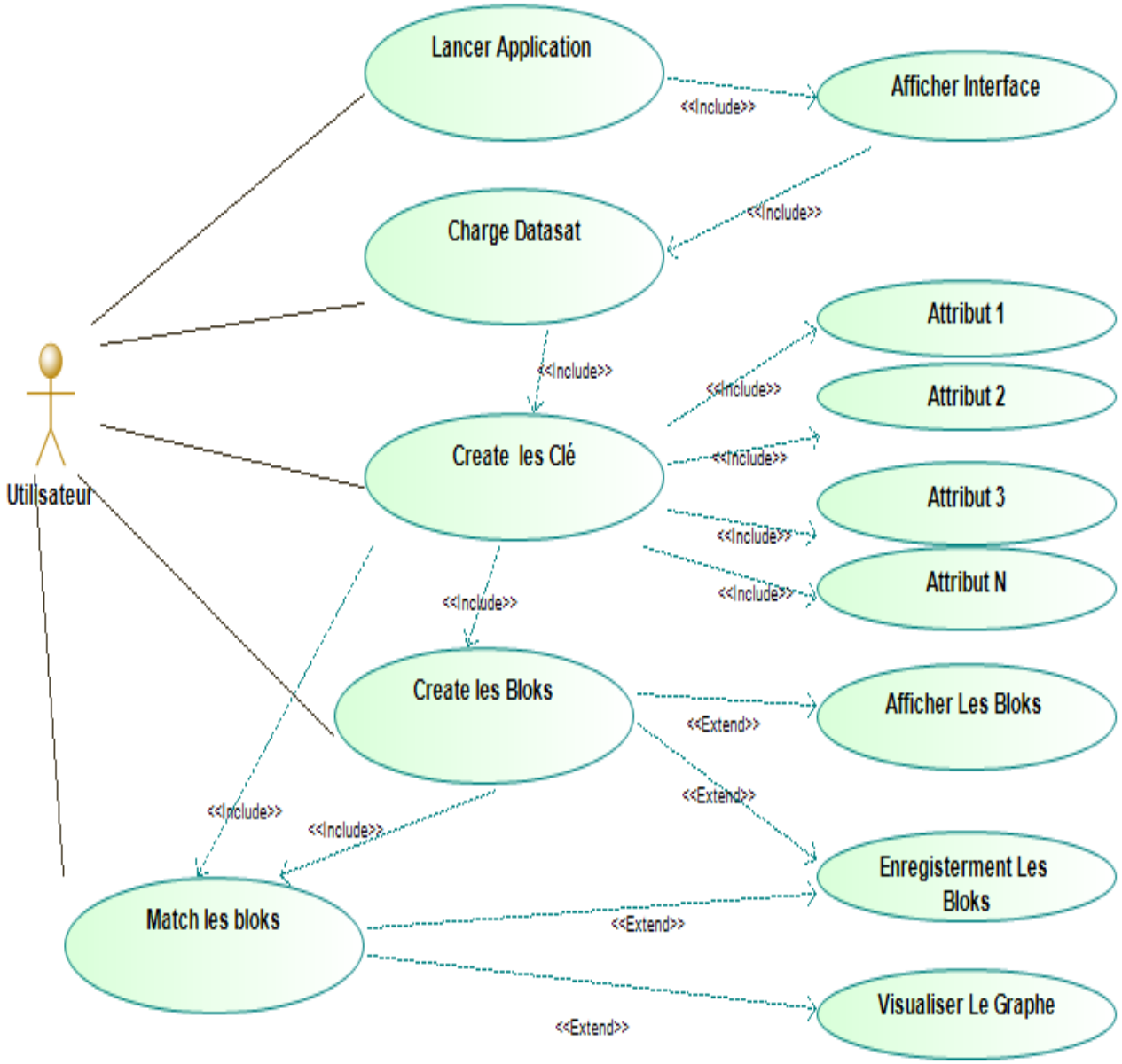


Figure 3.4 : Diagramme de cas d'utilisation d'acteur «Utilisateur»

III.6 Diagramme de séquence :

Pour mieux concrétiser les interactions entre les acteurs du système vis-à-vis de l'application et la base de données, nous sommes amenés à traduire nos scénarios en diagrammes de séquence.

Les diagrammes de séquence servent à illustrer les cas d'utilisation. Ils permettent de représenter des collaborations entre les objets selon un point de vue temporel, on y met l'accent sur la chronologie des envois des messages

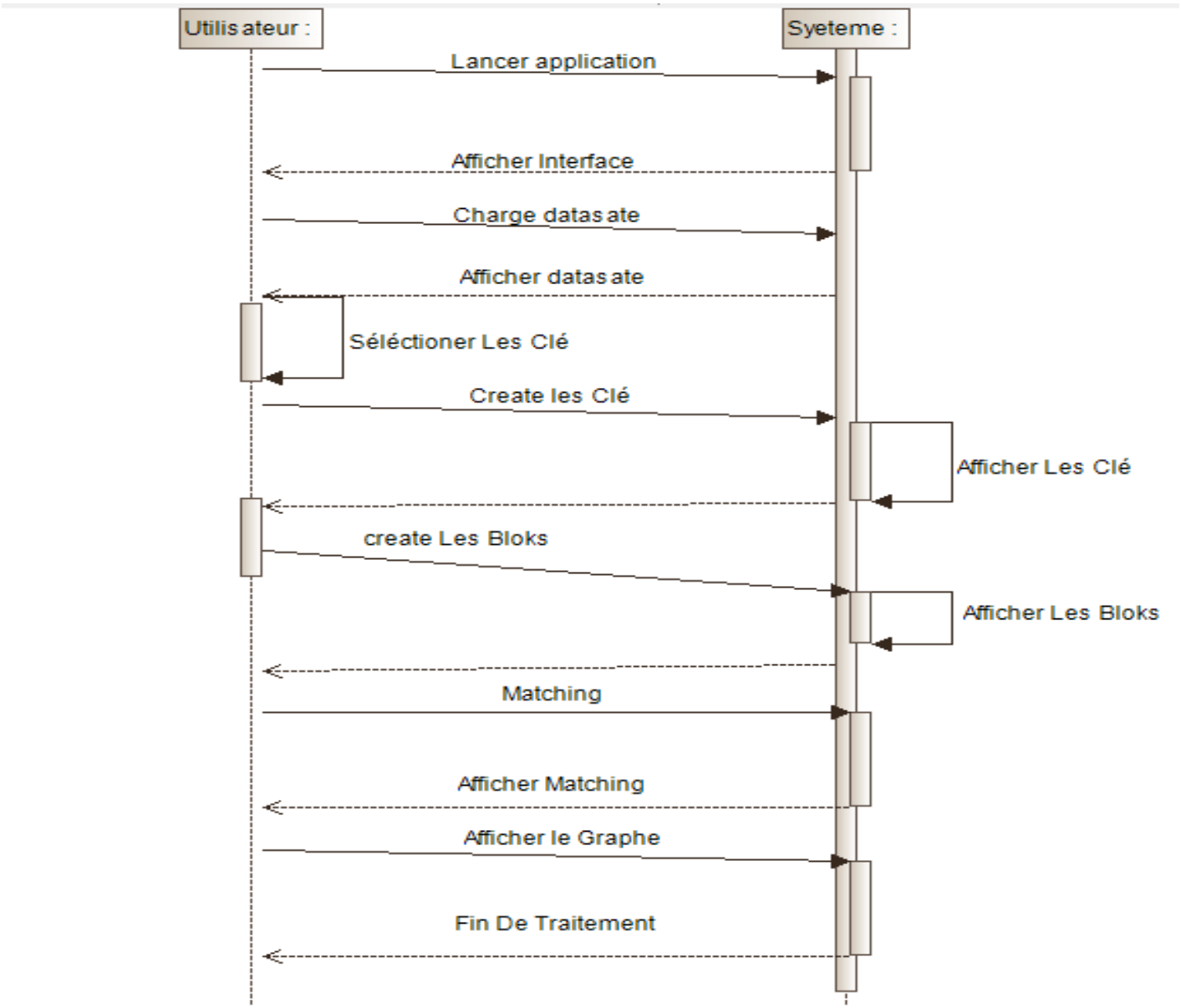


Figure 3.5 : Diagramme de Séquence d'acteur «Utilisateur»

III.7 Diagramme d'activités :

Dans la phase de conception, les diagrammes d'activités sont particulièrement adaptés à la description des cas d'utilisation. Plus précisément, ils viennent illustrer et consolider la description textuelle des cas d'utilisation. De plus, leur représentation sous forme d'organigrammes les rend facilement intelligibles et beaucoup plus accessibles que les diagrammes d'états-transitions. On parle généralement dans ce cas de modélisation de workflow. On se concentre ici sur les activités telles que les voient les acteurs qui collaborent avec le système dans le cadre d'un processus métier.

Nœuds d'activités :

De la gauche vers la droite, on trouve : le nœud représentant une action, qui est une variété de nœud exécutable, un nœud objet, un nœud de décision ou de fusion, un nœud de bifurcation ou d'union, un nœud initial, un nœud final et un nœud final de flot.

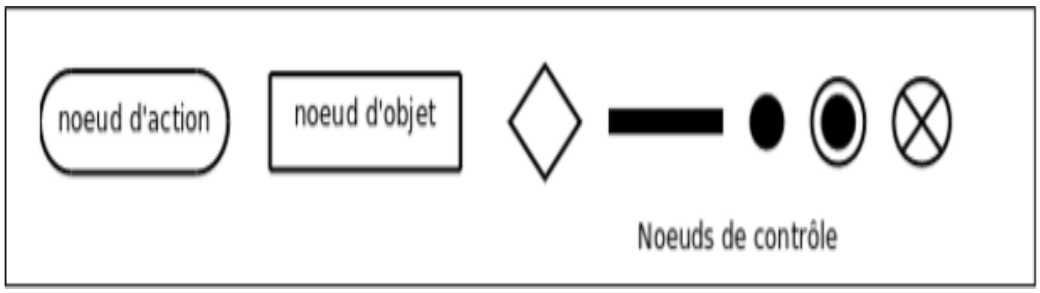


Figure 3.6 : Représentation graphique des nœuds d'activité

III.8 Diagramme d'activités « d'affectation des Utilisateur »:

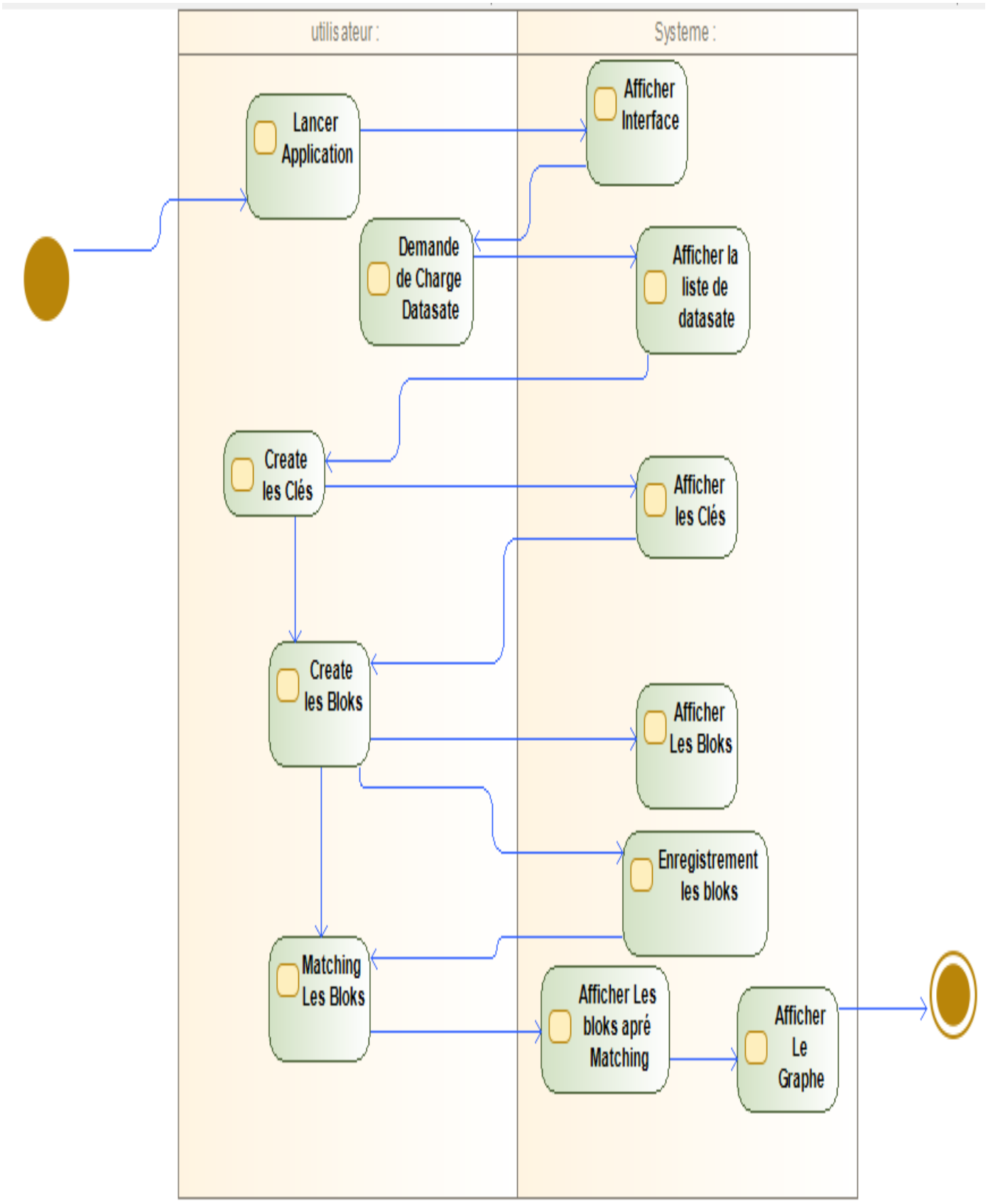


Figure 3.7 : Diagramme d'activités d'affectation des notes Utilisateur

III.9 Diagramme de classes :

Le diagramme de classe est une description statique du système focalisé sur le concept de classe et d'association. Une classe représente un ensemble d'objets qui possèdent des propriétés similaires et des comportements communs décrivant en terme d'attributs et d'opérations.

Une association consiste à présenter les liens entre les instances de classe. Durant cette section, nous allons présenter les diagrammes de classes entités à notre application.

III.9.1 Diagramme de classe « Utilisateur»:

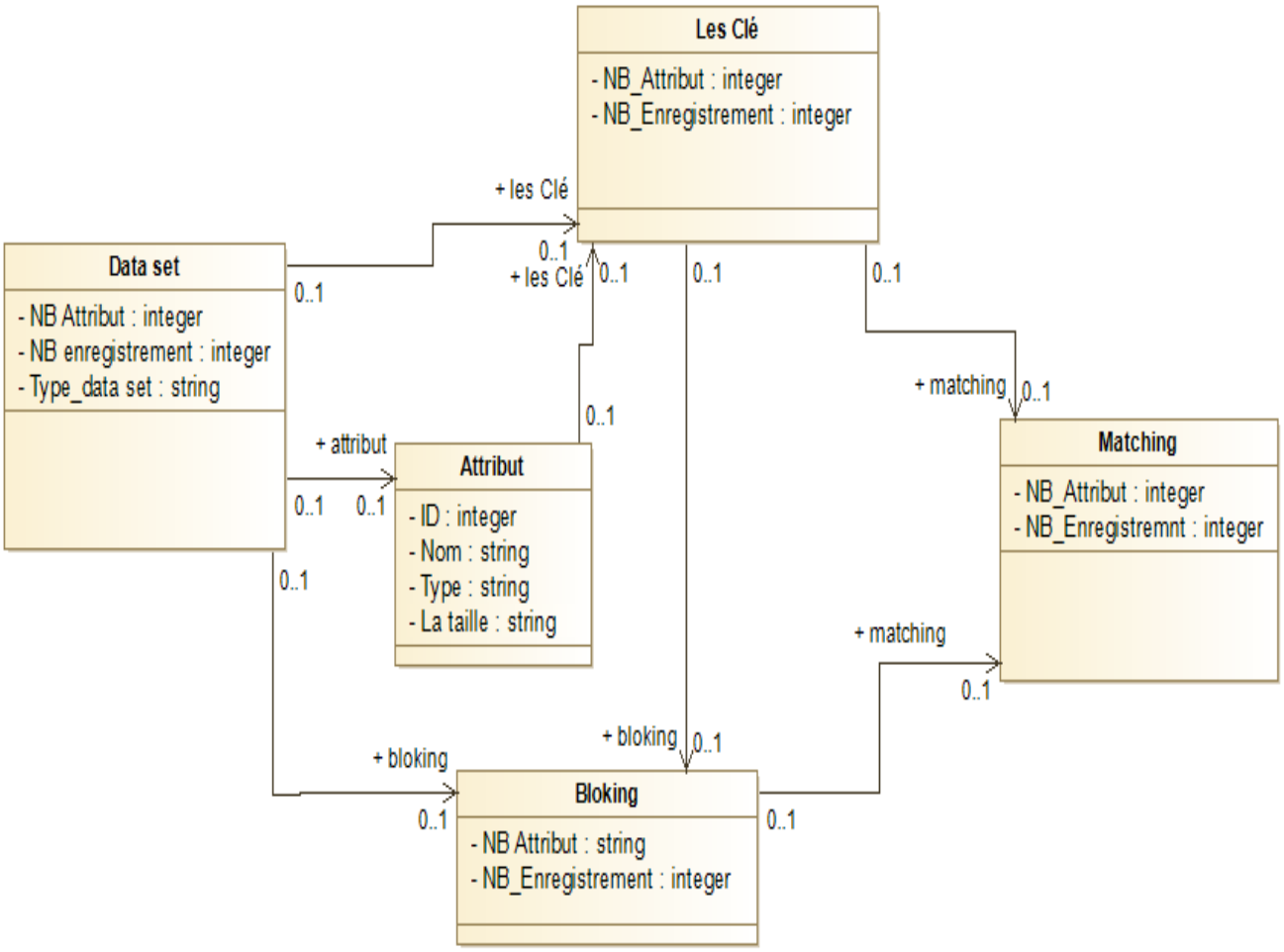
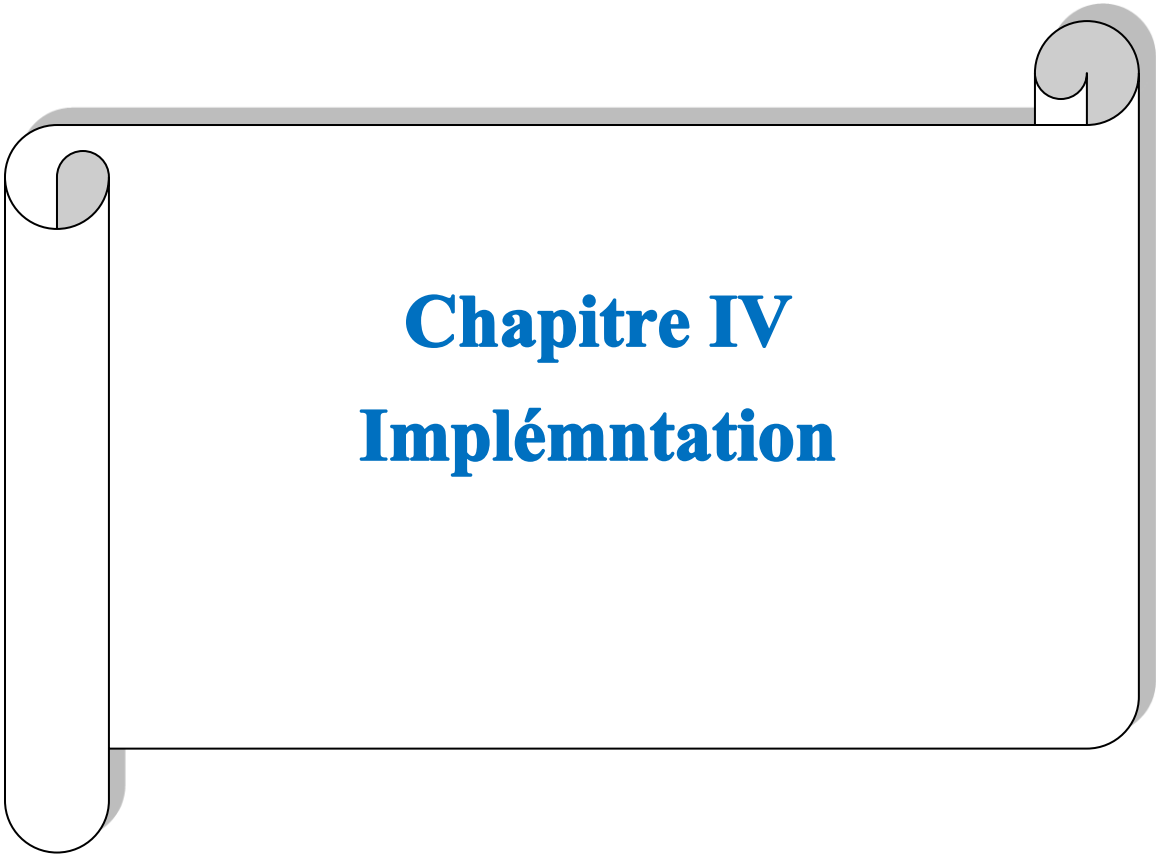


Figure 3.8 : Diagramme de classe « Utilisateur»

III.10 Conclusion:

Le logiciel de modélisation UML Visual Paradigme for UML est un bon outil pour réaliser des spécifications. Possédant de nombreuses fonctionnalités dans ses version payante, il possède une interface intuitive et est rapidement pris en main.



Chapitre IV
Implémntation

IV.4.1 Introduction:

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

IV.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

-Est La taille de data set 863

IV.4.3 languge utiliser :

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

a. JAVA est un langage multiplateformes qui permet aux concepteurs,Selon le principe : write once,run envery where d'écrire un code capable de fonctionner dans tous les environnements (quelque soit système d'exploitation).
b. 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).
c. JAVA est un langage orienté objets, simple qui réduit le risque d'erreurs et d'incohérence.
d. 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.
e. Un accès simplifié aux bases de données, soit à travers la passerelle JDBC-ODBC ou à travers un pilote JDBC spécifique au SGBD.
f. 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
g. Le développement avec java est gratuit.

IV.4.4 L'environnement de développement:

IV.4.4.1 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 [16].

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

a. 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 [16].
b. 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.

IV.4.4.2 JavaFX :

- 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 [17].

-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 [17].

IV.4.5 Les avantages :

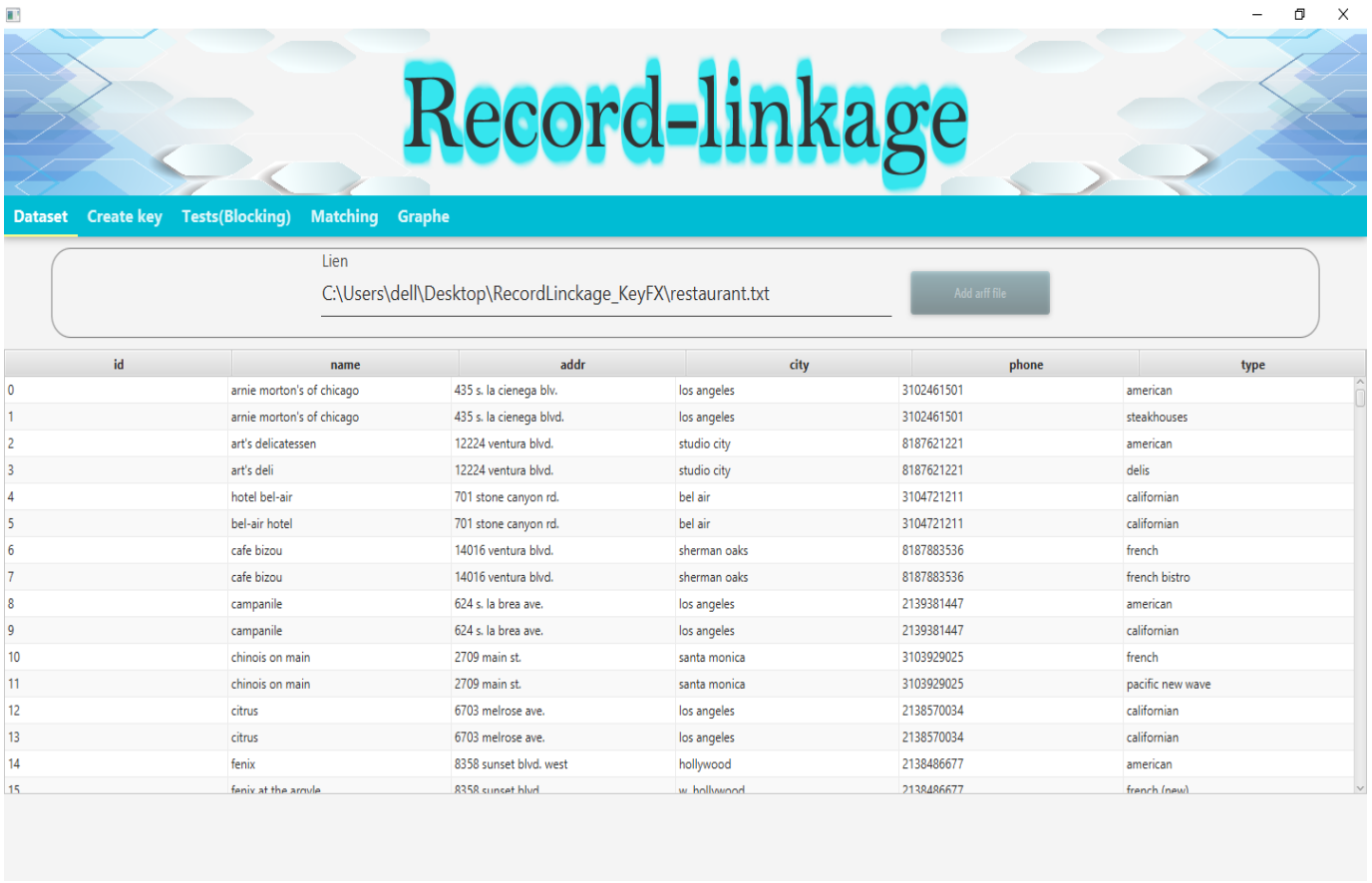
a. Basé sur Java (Java SE et ME).
b. Utilisable sur tous les écrans : navigateurs, mobile, TV, etc.
c. Open Source.
d. Déploiement sur navigateur et ordinateur de bureau "Desktop" sans modification.
e. Collaboration designers et développeurs. Possibilité d'intégrer des codes en Java et JavaFX
f. Moins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

IV.4.6 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
Data set sur laquelle nous voulons travailler

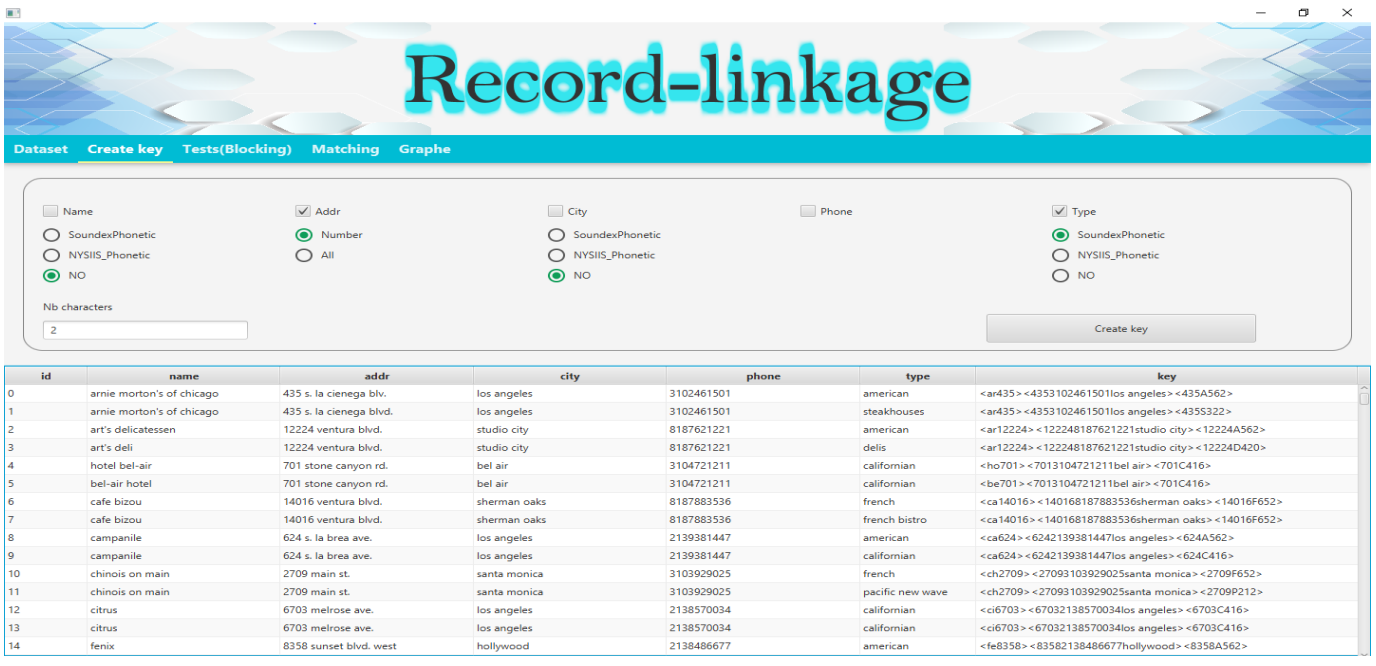
FIGURE 4.1 : Interface principale de notre application

FIGURE 4.2 : Sélectionnes fichier data set



id	name	addr	city	phone	type
0	arnie morton's of chicago	435 s. la cienega blv.	los angeles	3102461501	american
1	arnie morton's of chicago	435 s. la cienega blvd.	los angeles	3102461501	steakhouses
2	art's delicatessen	12224 ventura blvd.	studio city	8187621221	american
3	art's deli	12224 ventura blvd.	studio city	8187621221	delis
4	hotel bel-air	701 stone canyon rd.	bel air	3104721211	californian
5	bel-air hotel	701 stone canyon rd.	bel air	3104721211	californian
6	cafe bizou	14016 ventura blvd.	sherman oaks	8187883536	french
7	cafe bizou	14016 ventura blvd.	sherman oaks	8187883536	french bistro
8	campanile	624 s. la brea ave.	los angeles	2139381447	american
9	campanile	624 s. la brea ave.	los angeles	2139381447	californian
10	chinois on main	2709 main st.	santa monica	3103929025	french
11	chinois on main	2709 main st.	santa monica	3103929025	pacific new wave
12	citrus	6703 melrose ave.	los angeles	2138570034	californian
13	citrus	6703 melrose ave.	los angeles	2138570034	californian
14	fenix	8358 sunset blvd. west	hollywood	2138486677	american
15	fenix at the argyle	8358 sunset blvd.	w. hollywood	2138486677	french (new)

FIGURE 4.3 : Afficher la fichier data set



id	name	addr	city	phone	type	key
0	arnie morton's of chicago	435 s. la cienega blv.	los angeles	3102461501	american	<ar435><4353102461501los angeles><435A562>
1	arnie morton's of chicago	435 s. la cienega blvd.	los angeles	3102461501	steakhouses	<ar435><4353102461501los angeles><435S322>
2	art's delicatessen	12224 ventura blvd.	studio city	8187621221	american	<ar12224><12224818762122 1studio city><12224A562>
3	art's deli	12224 ventura blvd.	studio city	8187621221	delis	<ar12224><12224818762122 1studio city><12224D420>
4	hotel bel-air	701 stone canyon rd.	bel air	3104721211	californian	<ho701><7013104721211bel air><701C416>
5	bel-air hotel	701 stone canyon rd.	bel air	3104721211	californian	<be701><7013104721211bel air><701C416>
6	cafe bizou	14016 ventura blvd.	sherman oaks	8187883536	french	<ca14016><1401681878835 36sherman oaks><14016F652>
7	cafe bizou	14016 ventura blvd.	sherman oaks	8187883536	french bistro	<ca14016><1401681878835 36sherman oaks><14016F652>
8	campanile	624 s. la brea ave.	los angeles	2139381447	american	<ca624><624213938144 7los angeles><624A562>
9	campanile	624 s. la brea ave.	los angeles	2139381447	californian	<ca624><624213938144 7los angeles><624C416>
10	chinois on main	2709 main st.	santa monica	3103929025	french	<ch2709><2709310392902 5santa monica><2709F652>
11	chinois on main	2709 main st.	santa monica	3103929025	pacific new wave	<ch2709><2709310392902 5santa monica><2709P212>
12	citrus	6703 melrose ave.	los angeles	2138570034	californian	<ci6703><670321385700 34los angeles><6703C416>
13	citrus	6703 melrose ave.	los angeles	2138570034	californian	<ci6703><670321385700 34los angeles><6703C416>
14	fenix	8358 sunset blvd. west	hollywood	2138486677	american	<fe8358><8358213848667 7hollywood><8358A562>

FIGURE 4.4 : Choix les Clé

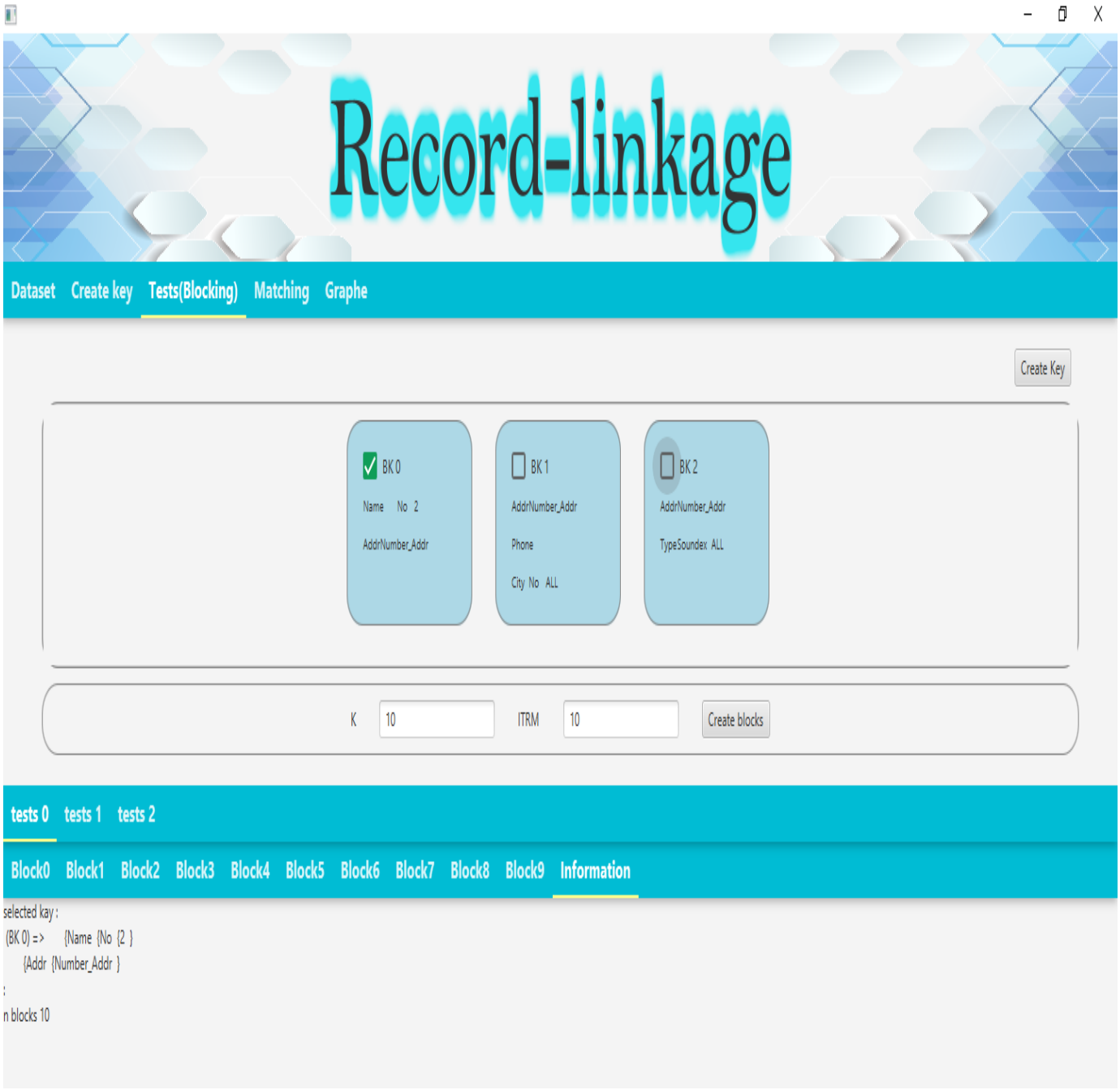


FIGURE 4.5 : Information des chaque tests

FIGURE 4.6 : Création Les Blocks

FIGURE 4.7 : information de détail matching

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

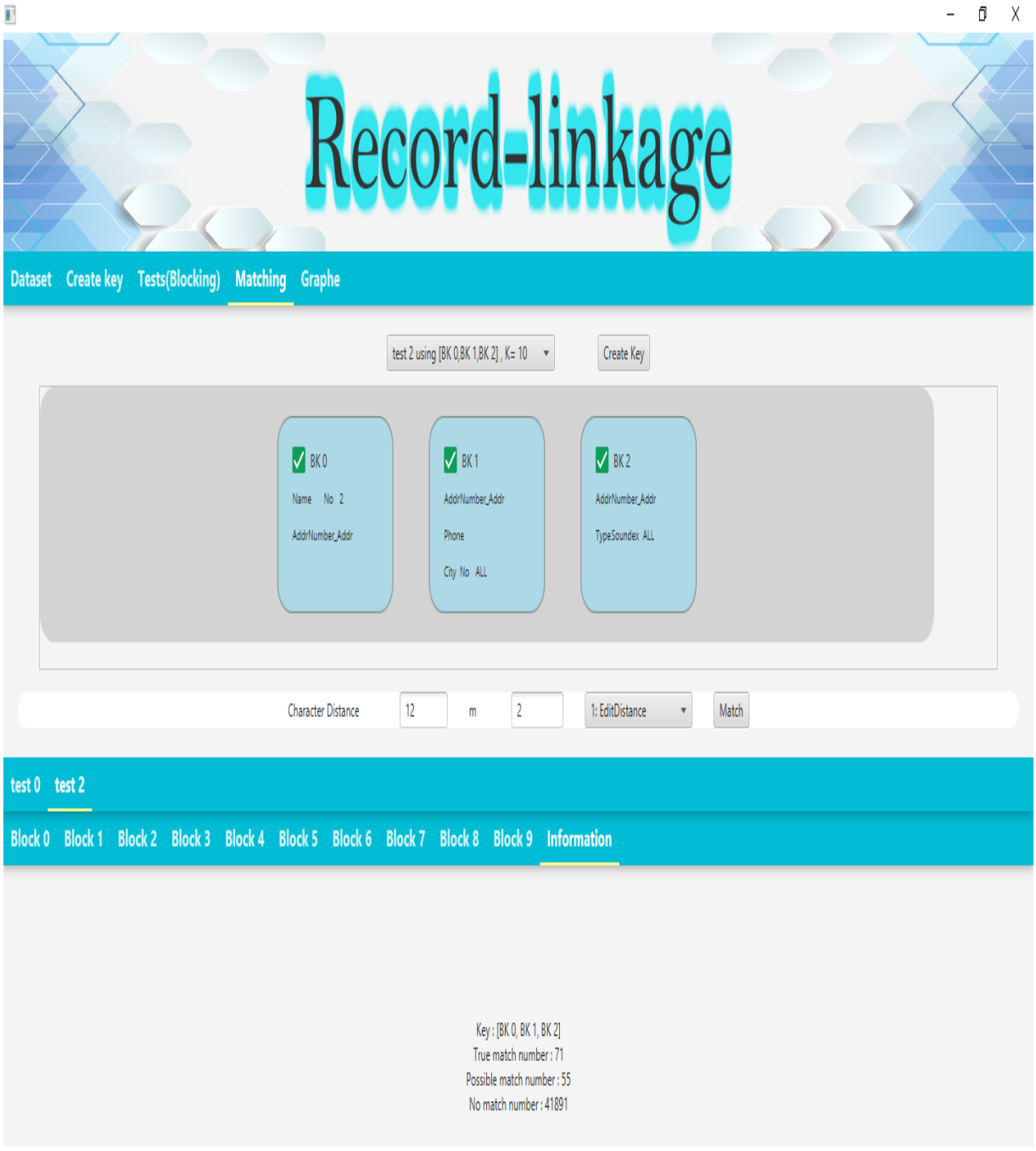


FIGURE 4.8 : Match Les Blocks

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

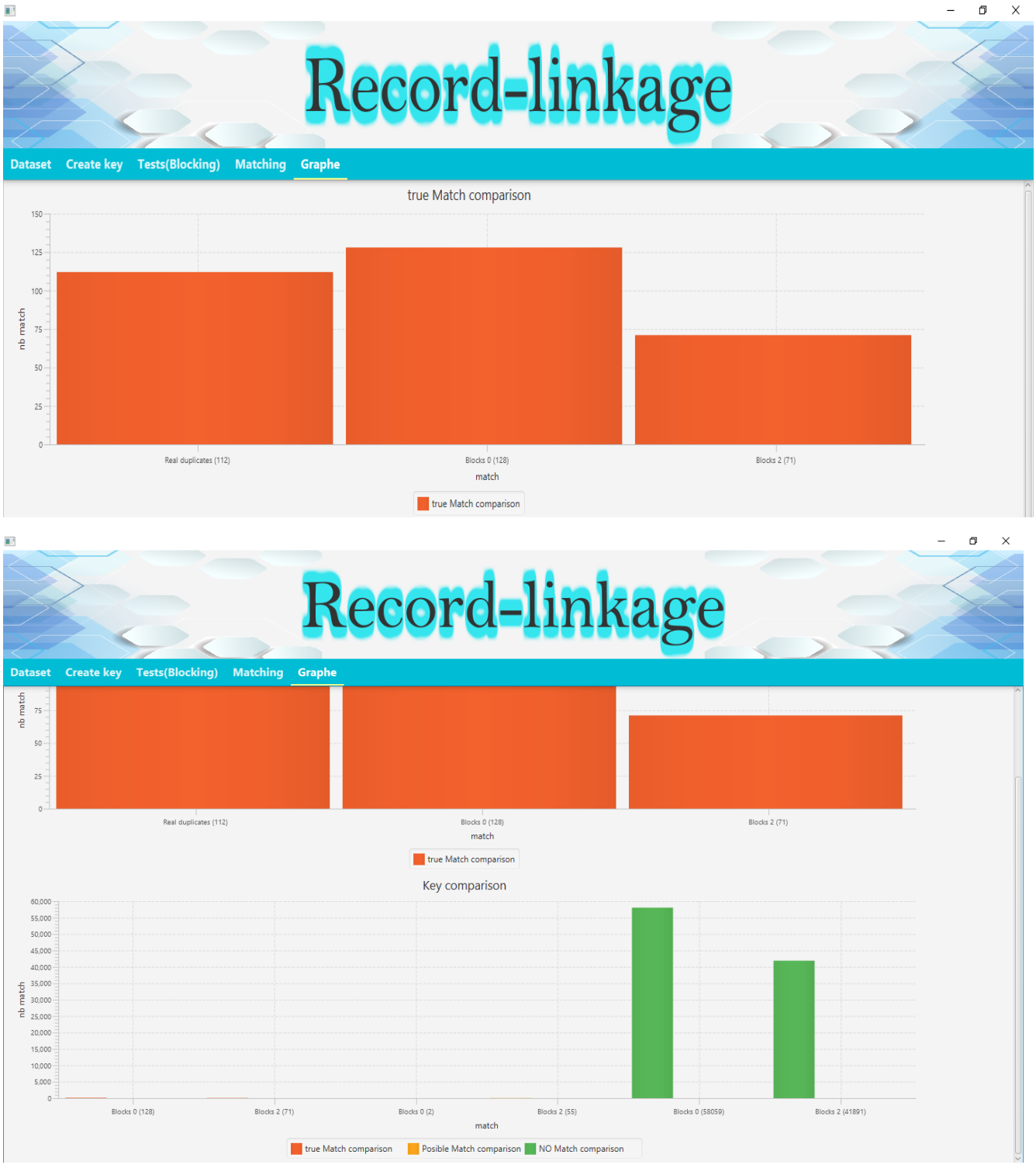


FIGURE 4.9 : Affiche les résultats du record (Graphe)

4.7 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

Notre étude est concentrée sur la sélection des attributs et la création des clés de blocage durant le processus de couplage d'enregistrements ou le record linkage (RL).

Le choix des clés de blocage durant l'étape de l'indexation ou blocage permet de amélioré la qualité de données en niveau chaque bloc obtenu.

L'étape de correspondance (matching) consiste à faire correspondre les paires d'enregistrements dans le même bloc et décider si elles représentent une correspondance vraie ou possible, cette décision a été faites en comparant les clés de blocage (BK) au lieu de comparer tous les attributs des data sets.

Sélection des attributs dans l'étape de correspondance (matching) pour crée des nouveau clés de blocage différents de celle utilisée dans l'étape d'indexation nous a permet d'obtenir des résultats avantages.

BIBLIOGRAPHIE :

[1] Mr Tabetabdelkrim, Quali de donne-Dtection de doublons et des dimi- laires,mmoire de Master,UniversitDjillali-Liabes de Sidi-Bel-Abbs,(2016/2017).

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

[3] Bassour Boumediene,AbbarRiadhWassim, Dtectionde doublons,Thse de master,Universit Djilali Liabes, Sid-Bel-Abess,(2015/2016). Franck Rgnier-Pcastaing, Michel Gabassi, Jacques Finet.

[4] Franck Rgnier-Pcastaing, Michel Gabassi, Jacques Finet, Enjeux et mthodes dela gestion des donnes,livre,(2008)

[5] JEMM research, DES DONNES QUALIT :Exploitez le capital de votre organisa-tion ,livre blanc,(janvier 2008).

[6] Laure Berti-quille, Qualit des donnes,Matre de Confrences

[7] Le bulletin technique TB00017, Donnees De Reference Sur La Deduplication De Donnees,Livre Blanc. A Propos De Quantum, (Juillet 2014).

[8] Z. Bahmani, L. Bertossi, N. Vasiloglou, ERBlox: combining matching dependen cies with machine learning for entity resolution, Int. J. Approx. Reason. 83(2017) 118–141.

[9] M. Bilenko, B. Kamath, R.J. Mooney, Adaptive blocking: learning to scale up record linkage, in: Proceedings of the Sixth International Conference on Data Mining, ICDM, IEEE, 2006, pp. 87–96.

[9] Aizawa. A et Oyama. K. A fast linkage detection scheme formultisource information integration. Dans In null, page 30–39. IEEE. Matthew A.Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. Journal of the American Statistical Association, 84(406) :414–420,

[9] Ramadan B et Christen P. Unsupervised blocking key selection for real-time entity resolution. Dans Pacific-Asia Conference on Knowledge Discovery and Data Mining, page 574–585. Springer.

[9] Bala.J, Huang. J, Vafaie. H, et DeJong. K et Wechsler. H. Hybrid learning using genetic algorithms and decision trees for pattern classification. in IJCAI, 1 :719–724.

[10] Laure Berti-Equille. Qualité des données. Techniques de l'ingénieur. Informatique, 2006.

[10] Bassour Boumediene et Abbar Riadh Wassim. Dtectionde doublons.

[10] Thése de master Universit Djilali Liabes ,Sid Bel Abess, 2015/2016.

[11] Muro C, Escobedo Rand Spector L, et Coppinger R. Wolf-pack (Canis lupus) hunting strategies emerge from simple rules in computational simulations. Behav Processl, volume 88.

[11] Peter Christen. A comparison of personal name matching : Techniques and practical issues. Dans Data Mining Workshops, *2006*. ICDM Workshops *2006*. Sixth IEEE International Conference on, page 290–294. IEEE,

[12] Peter Christen. A survey of indexing techniques for scalable record linkage and deduplication. IEEE transactions on knowledge and data engineering, 24(9) :,1537–1555,, b.
Peter Christen. Towards parameter-free blocking for scalable record linkage.
[13] Nascimento D.C, Pires C.E.S., et Mestre D.G. Exploiter la cooccurrence de bloc pour contrôler la taille des blocs pour la résolution d'entité. Knowledge and Information Systems, *62*(*1*), *359- 400*, 62(1) : 359–400.
[14] Emary E, Zawbaa H.M, Ghany K.K.A., et A.E.and Pârv B Hassanien. Firefly optimization algorithm for feature selection. Dans Actes de la *7*e Conférence des Balkans sur l'informatique, page 26. ACM.
[15] Ahmed.K Elmagarmid, Panagiotis.G Ipeirotis, et Vassilios S Verykios. Duplicate record detection : A survey. IEEE Transactions on knowledge and data engineering, 19(1) :1–16,
[15] Fleuret F. Fast binary feature selection with conditional mutual information. Journal of Machine learning research, 5(novembre) : 1531–1555 .
[16] Chandrashekar G et Sahin F. An introduction to variable and feature selection, guyon, isabelle and elisseeff, andré,. Journal of machine learning research, 3(1) :16–28. numéro : mars, pages : 1157–1182, année : 2003. Informatique et génie électrique,.
[17] TN Gadd. Phonix : The algorithm. Program, 24(4) :363–366,. Köpcke H, Thor A, et Rahm E. Evaluation of entity resolution approaches on real-world match problems. Proceedings of the VLDB Endowment, 3(1-2) :484–493.
[17] Köpcke H, Thor A, et Rahm E. Learningbased approaches for matching web data entities. IEEE Internet Computing, 14(4)

:23–31, b. Mauricio A Hernàndez et Salvatore J Stolfo. The merge/purge problem for large databases. Dans ACM Sigmod Record, volume 24, page 127–138. ACM.
[17] Benkhaled H.N, Berrabah D, et Boufares F. A novel approach to improve the record linkage process. Dans En *2019* IEEE *6*th International Conference on Control, Decision and Information Technologies, page 1504–1509. IEEE.
[17] David Holmes et M.Catherine McCabe. Improving precision and recall for soundex retrieval. Dans Information Technology : Coding and Computing, *2002*. Proceedings. International Conference on, page 22–26. IEEE.
[17] Guyon I et Elisseeff A. An introduction to variable and feature selection. Journal of Machine learning research, 3(mars) :1157–1182. Shao J et Wang Q. Active blocking scheme learning for entity resolution. Dans Pacific-Asia Conference on Knowledge Discovery and Data Mining, page 350–362, Cham. Springer. Jamm. DES DONNES QUALIT .
[17] organisation. livre blanc, janvier 2008. Gravano L, Ipeirotis P.G, Jagadish H.V., Koudas N, Muthukrishnan S, et Srivastava D. Approximate string joins in a database (almost) for free. VLDB, 1 :491–500, a. Pipino L.L, Lee Y.W, Wang, et R.Y. Data quality assessment. Communications of the ACM, 45(4) :211–218.
[17] Alian M, Awajan A, et Ramadan B. Unsupervised learning blocking keys technique for indexing arabic entity resolution. International Journal of Speech Technology, 22(3) :621–628,
[17] Bilenko M, Kamath B, et Mooney R.J. Adaptive blocking : Learning to scale up record linkage. in sixth international conference on data mining. Dans Sixth International Conference

on Data Mining (ICDM'*06*, page 87–96. IEEE, b. MichelsonMet Knoblock C.A. Learning blocking schemes for record linkage. AAAI, 6 :440–445.Hernández M.A et Stolfo S.J. The merge/purge problem for large databases. Dans ACM Sigmod Record, 24 :127–138.Andrew McCallum, Kamal Nigam, et Lyle H Ungar. Efficient clustering of high-dimensional data sets with application to reference matching. Dans Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, page 169–178. ACM.

[18] Dif N, Attaoui M, et Elberrichi Z. Gene selection for microarray data classification using hybrid meta-heuristics. Dans International Symposium on Modelling and Implementation of Complex Systems, page 119–132. Springer.

[18] Dif N et Elberrichi Z. An enhanced recursive firefly algorithm for informative gene selection. International Journal of Swarm Intelligence Research (IJSIR, 10(2) :21–33.

[18] Jona J.et Nagaveni et N. Ant-cuckoo colony optimization for feature selection in digital mammogram. Journal pakistanais des sciences biologiques, 17(2) :266.

[18] Abdelkrim OUHAB, Mimoun MALKI, Djamel BERRABAH, et Faouzi BOUFARES. An unsupervised entity resolution framework for english and arabic datasets. International Journal of Strategic Information Technology and Applications (IJSITA, 8(4) :16–29,

[18] Ivan P.Fellegi et Alan B.Sunter. A theory for record linkage. Journal of the American Statistical Association, 64(328) :,1183–1210,

[18] Franck Rgnier-Pcastaing, Michel Gabassi, et Jacques Finet. Enjeux et mthodes de la gestion des donnes. Paperback, 2008.

[18] Kalsi S, Kaur H, et Chang V. Dna cryptography and deep learning using genetic algorithm with nw algorithm for key generation. Journal of medical systems, 42(1) :17. Tobias Vogel et Felix Naumann. Automatic blocking key selection for duplicate detection based on unigram combinations. Dans Proceedings of the International Workshop on Quality in Databases (QDB.Yang Y, Zheng X, Guo W, Liu X, et Chang V. Privacy-preserving smart iot-based healthcare big data storage and self-adaptive access control system. Information sciences, 479 :567–592.

[18] Su Yan, Dongwon Lee, Min-Yen Kan, et Lee C Giles. Adaptive sorted neighborhood methods for efficient record linkage. Dans Proceedings of the *7*th ACM/IEEE-CS joint conference on Digital libraries, page 185–194. ACM.

Public License :

[1].https://www.futurasciences.com/tech/definitions/internet java-485/

[2]. https ://bit.ly/2I9RGeg

[3]. https ://www.techno-science.net/definition/5346.html

[4]. https ://www.cs.waikato.ac.nz/ml/weka/

[5]https://www.ibm.com/docs/fr/rationalsoftarch/9.5?topic=diagrams-uml-models

[6] https://www.techno-science.net/definition/764.html

[7] https://laurent-audibert.developpez.com/Cours-UML/?page=diagramme-cas-utilisation

[8] https://laurent-audibert.developpez.com/Cours-UML/?page=diagrammes-interaction

[9] https://laurent-audibert.developpez.com/Cours-

UML/?page=diagramme-classes

[10] https://laurent-audibert.developpez.com/Cours-UML/?page=diagramme-activity

[11] https://dspace.univouargla.dz/jspui/bitstream/123456789/1617/1/Master-Ahfouda-Habbi.pdf

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

Table des figure

Liste des tableaux