Juin 2024

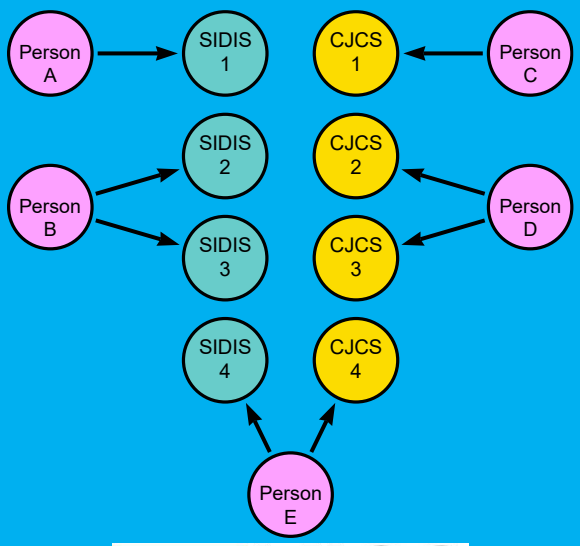
60

Le développement d'une base de données historique intégrée pour l'étude de la récidive et des carrières criminelles.

Rapport de recherche.

Patrick Jeuniaux, Benjamin Mine, Luc Robert, Eric Maes,

Michaël Vande Velde



Nationaal Instituut voor Criminalistiek en Criminologie

Institut National de Criminalistique et de Criminologie

Projet financé par la Politique scientifique fédérale (**BELSPO**)

Contrat **B2/202/P2/IIHA** https://incc.fgov.be/IIHA

Le projet IIHA à la base de la recherche décrite dans ce rapport s'est déroulé sur une période de deux ans (2021-2023) et a bénéficié du soutien financier de la politique scientifique fédérale belge (BELSPO) dans le cadre du Pilier 2 de la « Science du patrimoine » du programme de recherche « BRAIN-be 2.0 ». Le Pilier 2 est destiné à soutenir la conservation des données fédérales afin de favoriser leur exploitation scientifique, historique et statistique[1].

Nous remercions nos collègues du Service Public Fédéral Justice pour l'aide précieuse qu'ils ont apportée tout au long de cette recherche, que ce soit par l'accès aux données nécessaires à l'analyse ou les éclairages apportés par rapport à la signification de ces données.

Nous remercions également les collaborateurs et collègues avec qui nous avons pu discuter des sujets décrits dans ce travail : Julien Dupont, Pierre Hauweele, Alain Hertz, Philippe Huynen, Hadrien Mélot, Thierry Morre, Michaël Person, Marco Saerens, Samuel Waltener.

Enfin, nous remercions vivement Tom Geudens (Neo4j) pour avoir apporté son expertise en matière de modélisation et d'intégration de données dans Neo4j, tant par ses conseils concernant la stratégie à suivre, qu'en termes de programmation des requêtes en Cypher pour mettre cette stratégie en œuvre.

Malgré les nombreux cycles de relecture et échanges sur le sujet, les auteurs restent responsables des erreurs, approximations ou omissions qui demeureraient dans ce document.

---

[1] Programme Brain-be 2.0 (Belgian Research Action through Interdisciplinary Networks) https://www.belspo.be/belspo/brain2-be/index_fr.stm

Table des matières

Liste des figures

Liste des tableaux

1. Introduction

Dans le cadre de ses activités de recherche en criminologie, l'Institut National de Criminalistique et de Criminologie (INCC) sollicite auprès de ses partenaires institutionnels l'accès à différents types de données, dont des données digitales relatives à l'administration de la justice pénale.

Ces données digitales sont riches en informations et couvrent des décennies d'enregistrements. Le projet IIHA repose sur deux ensembles de données historiques provenant du Casier judiciaire central (CJCS) lequel traite des données relatives notamment aux condamnations, suspensions et internements et de l'ancienne base de données SIDIS-greffe (SIDIS) qui enregistre des données relatives aux détentions.

Ce sont des données historiques au sens où elles ont été extraites à un moment particulier dans le temps. En l'occurrence, les données disponibles en provenance de SIDIS-greffe ont été extraites en 2014, c'est-à-dire l'année où l'usage de cette base de données a pris fin[2]. Quant à celles en provenance de CJCS, elles ont été extraites en 2020, et nous n'avons pu disposer d'une extraction plus récente depuis.

Dans le passé, les chercheurs de l'INCC ont exploité les données de CJCS[3] ainsi que les données de SIDIS-greffe[4]. Leur tâche a cependant été rendue ardue pour trois types de raison.

Raison 1 : « faiblesse de la documentation ». La documentation concernant ces deux bases de données était parcellaire, rendant leur compréhension difficile.

Raison 2 : « données brutes ». Les données disponibles existaient sous une forme brute, à l'état de fichiers d'extraction, qu'il fallait réorganiser pour les besoins de chaque analyse.

Raison 3 : « pas d'identifiant unique de la personne ». L'analyse simultanée des données de condamnation et de détention était rendue difficile par l'absence d'un identifiant unique permettant de reconnaître ces personnes au sein de ces deux bases de données.

Un tel état des choses n'était pas propice à une exploitation pérenne des données basée à la fois sur les données de condamnation et les données de détention.

Le travail de réflexion au sein de l'INCC sur les conditions d'articulation des données n'est pas neuf (Mine et Vanneste, 2011; Vanneste, 2012) et un travail préliminaire d'exploration des données fut déjà entrepris (De Blander et al. 2019), notamment pour envisager l'usage simultané des données de condamnation et de détention (Raison 3 « pas d'identifiant unique de la personne »). Ce travail fut ensuite étendu dans le cadre du projet de recherche « FAR » (Jeuniaux, Mine, et Detry, 2022), via

---

[2] Étant donné qu'elle fut cette année-là remplacée par une nouvelle base de données : SIDIS-suite.

[3] « Désister ou persister ? » : La première étude nationale sur la récidive en Belgique. https://incc.fgov.be/stoppen-of-doorgaan-het-eerste-nationale-onderzoek-naar-recidive-in-belgie-desister-ou-persister-la

[4] Strafuitvoering (exécution des peines). https://incc.fgov.be/strafuitvoering

l'usage de nouvelles méthodes d'intégration des données, ainsi que par le rassemblement des données brutes au sein d'une nouvelle base de données vouée à leur exploitation (Raison 2 « données brutes »).

Cependant, la connaissance des deux bases de données sources (CJCS et SIDIS) demeurait peu accessible en raison de l'absence de documentation spécifiquement dédiée à la description de ces bases de données (Raison 1 « faiblesse de la documentation »).

Malgré les progrès effectués dans le cadre de cette dernière recherche, la situation demeurait peu propice à une exploitation pérenne et combinée de ces deux sources de données. Le projet de recherche IIHA[5] entend précisément corriger ces manquements, préserver au mieux ce patrimoine numérique pour les recherches criminologiques futures et en faciliter l'exploitation afin d'étudier la récidive et les carrières criminelles, c'est-à-dire la caractérisation de la séquence longitudinale des crimes commis par les délinquants individuels (Blumstein et al. 1986).

Le projet IIHA poursuivait plus particulièrement quatre objectifs principaux.

1) Documenter les deux bases de données d'intérêt (CJCS et SIDIS)[6].

2) Développer une Base de Données Historiques Intégrée (IHD[7]) pour stocker, intégrer et exploiter les jeux de données extraits de chacune de ces deux bases de données d'intérêt.

3) Exploiter la base de données intégrée afin d'effectuer des analyses statistiques sur la récidive et les carrières criminelles dans le cadre d'études criminologiques.

4) Élaborer un prototype de « moniteur de la récidive » qui permet de mesurer et suivre la récidive à partir des données dynamiques de CJCS[8].

Ces différentes activités ont dans leur réalisation bénéficié l'une de l'autre, et ont fait l'objet de rapports ou d'articles spécifiquement dédiés, à l'exception du second objectif sur la Base de Données Historiques Intégrée (IHD). C'est de l'IHD dont il est question dans le présent rapport de recherche.

Nous rendons ici compte du travail effectué, en mettant l'accent sur les innovations méthodologiques apportées dans cette nouvelle version de la base de données, par rapport à celle développée dans le cadre du projet FAR. Nous expliquons l'utilité de cette approche ainsi que certaines de ses limites. À titre illustratif du potentiel de l'IHD, nous présentons des résultats concernant l'exécution de la procédure d'intégration des enregistrements, ainsi que des résultats relatifs aux personnes ainsi identifiées. Ce court rapport ne donne pas toutes les informations nécessaires pour comprendre dans le détail la méthodologie employée, ni toutes les informations se rapportant aux résultats. Ces informations feront l'objet de publications ultérieures.

---

[5] Voir https://incc.fgov.be/IIHA

[6] Voir Huynen, Jeuniaux, et al. (2024) pour CJCS et Maes, Mine et al. (2024) pour SIDIS.

[7] Nous nous contentons ici de recourir à l'acronyme de sa dénomination en anglais : « Integrated Historical Database » (IHD).

[8] Dynamiques au sens où il s'agit des données disponibles en temps réel via CJCS, et non pas de données contenues dans une extraction obtenue à un temps t. Voir Huynen, Mine, et al. (2024).

2. Méthodologie

2.1. Les objectifs

Aux fins de ce projet, nous disposons de deux ensembles de données. Le premier ensemble consiste en une extraction complète des données de condamnation issues de la base de données du Casier Judiciaire Central (CJCS) jusqu'à octobre 2020. Le second ensemble consiste en une extraction complète des données d'emprisonnement issues de la base de données SIDIS-greffe (SIDIS) des Établissements pénitentiaires de 1974 à septembre 2014, date de la migration vers la nouvelle application SIDIS-suite.

Le premier objectif de ce projet est de préserver ces données dans le temps et d'en faciliter l'exploitation pour les recherches futures. Pour cela une nouvelle base de données nécessite d'être créée afin d'accueillir ces données et de les rendre accessibles. Autrement dit, il s'agira de créer un entrepôt de données (« datawarehouse »).

Le second objectif de ce projet est de faciliter l'exploitation conjointe de ces deux ensembles de données au niveau de la personne (une même personne pouvant ou non se trouver dans chacune de ces deux bases de données). De cette manière il sera possible de soutenir l'étude de la récidive et des carrières criminelles qui utilise à la fois des données de condamnation et des données de détention. Comme ce sont des données provenant de sources distinctes pour lesquelles des identifiants uniques de la personne ne sont pas toujours disponibles, il s'agit de créer de nouveaux identifiants à partir d'une procédure à définir. Autrement dit, il s'agit de définir et appliquer une procédure d'intégration des données au niveau de la personne, et ainsi permettre l'exploitation de ces données intégrées pour étudier la récidive et les carrières criminelles.

La réalisation de ces deux objectifs se fera via le développement d'une base de données historique intégrée (IHD). L'IHD est donc à la fois un entrepôt de données et un outil pour intégrer et exploiter les données.

2.2. Les procédures d'intégration

L'intégration des données consiste à identifier dans un jeu de données, ou plusieurs jeux de données, les enregistrements qui se réfèrent à la même entité. Dans le cas qui nous occupe, il s'agit d'identifier les enregistrements qui se réfèrent à la même personne.

Initiées par les travaux de Dunn (1946), Fellegi et Sunter (1969) et Newcombe et al. (1959), de telles procédures sont aujourd'hui connues dans la littérature anglophone sous différentes appellations parmi lesquelles « record linkage », « data matching », « data linkage », « entity resolution ».

Il existe deux grands types de méthode d'intégration des données : les méthodes déterministes et les méthodes probabilistes. Quel que soit le type de méthode, elles se basent toutes sur une première étape de pré-traitement des données, qui cherchent à standardiser les données, afin de pouvoir les comparer au mieux, puis les intégrer. Les méthodes probabilistes calculent automatiquement les probabilités que les enregistrements soient liés à la même entité, sur la base des données. Les

méthodes déterministes, quant à elles, s'appuient sur un ensemble de règles, définies par l'analyste, qui décident, quand deux enregistrements ont suffisamment d'informations similaires ou en commun, qu'ils sont en fait associés à la même entité. Les méthodes déterministes fonctionnent suffisamment bien quand les enregistrements disposent de suffisamment d'identifiants de bonne qualité en commun (par exemple, le prénom, le nom, la date de naissance). Dans le présent travail, c'est une méthode déterministe qui a été mise au point.

2.3. Les bases de données en graphe

Dans le cadre de cette recherche, nous avons développé l'IHD dans Neo4j[9], une base de données en graphe.

Toutes les bases de données en graphe utilisent la notion de graphe, c'est-à-dire un ensemble de nœuds reliés par des relations, les nœuds pouvant représenter des entités du monde réel (e.g., une personne) et les relations, les relations entre ces entités (e.g., « est l'ami de »). Cette idée toute simple permet de représenter de nombreux phénomènes d'intérêt, et de tirer profit des concepts de la théorie des graphes et des algorithmes qui y sont associés. En particulier, Neo4j met à disposition de l'utilisateur des algorithmes utiles pour la résolution d'entités[10], et nous avons utilisé certains d'entre eux dans le cadre de ce travail.

Une propriété particulière de Neo4j est le caractère flexible et itératif de son usage (Robinson, Webber, et Eifrem 2015). Le fait que l'information soit représentée par des nœuds distincts, implique que chaque nœud puisse contenir une information différente de celle contenue par les autres nœuds. Par exemple, un nœud représentant une personne peut avoir un prénom et un nom, mais être dépourvu de date de naissance, et un autre nœud représentant une autre personne peut avoir un prénom, un nom mais également une date de naissance. Il en est de même des relations. Autrement dit, il n'y a pas dans Neo4j de définition stricte de la manière dont l'information est représentée qui s'impose à tous les éléments du graphe. Par conséquent, l'analyste chargé de modéliser l'information dans la base de données est libre de représenter des cas particuliers d'une manière qui soit très différente du cas général et aussi d'adapter la manière de représenter les choses en cours de route, au gré des besoins de son analyse. Une telle situation permet un démarrage rapide dans le développement de la base de données, et est aussi utile pour gérer des données disparates et évolutives.

Au point de départ du projet IIHA, la connaissance des données était déjà bien avancée (voir Jeuniaux et al. 2022) mais encore incomplète et susceptible d'évoluer. Par ailleurs, les documents produits au sujet de CJCS (Huynen, Jeuniaux, et al., 2024) et de SIDIS (Maes, Mine et al., 2024) n'étaient pas encore disponibles à ce moment-là. Et quand bien même ils l'auraient été, ceux-ci ne couvrent pas dans le détail la totalité des tables présentes dans les extractions reçues. Or il s'agissait de créer un entrepôt de données avec la totalité des données disponibles. Dans un tel contexte, une certaine flexibilité dans la manière de représenter les données demeurait nécessaire.

[9] https://neo4j.com

[10] https://neo4j.com/blog/graph-data-science-use-cases-entity-resolution/

En outre, à mesure que progressait notre compréhension du problème de l'intégration des données, nous avons mis au point plusieurs méthodologies pour réaliser cette intégration, ce qui a mené à la création successive dans le graphe de nœuds et relations de types nouveaux, sans que cela déstabilise le système. Neo4j est en effet conçue pour pouvoir modéliser les données de manière itérative.

Ensuite, les bases de données en graphe natives telles que Neo4j disposent de pointeurs qui donnent accès directement aux données sans passer par des jeux d'index, ce qui autorise des gains de rapidité lorsqu'il s'agit de devoir traverser le graphe un grand nombre de fois (Cheng et al. 2019; Robinson et al. 2015). C'est une propriété que nous utilisons dans le cadre de ce travail, dans la mesure où la méthode d'intégration qui a été mise au point explore le graphe des enregistrements, or ceux-ci sont liés les uns aux autres de très nombreuses manières.

Enfin, outre les algorithmes utiles pour la résolution d'entités, Neo4j met à disposition des outils de gestion et de science de données, en constant développement. Nous espérons pouvoir utiliser ce potentiel dans le futur.

2.4. Les données

Les données qui sont appelées à être stockées et exploitées dans l'IHD proviennent de deux sources : CJCS et SIDIS (voir Tableau 1). Ce sont des données « historiques » dans le sens où ce sont des extractions de données effectuées à des moments particuliers dans le temps.

Les données de CJCS ont été extraites de leur système informatique le 23 octobre 2020 et existent sous la forme de 64 fichiers CSV, tandis que les données de SIDIS ont été extraites le 10 octobre 2014 et existent sous la forme de 47 fichiers CSV.

Des données de CJCS plus récentes que celles de 2020 pourraient être obtenues dans le futur, mais en l'occurrence, ce sont les seules dont nous disposons pour le moment. Les données de SIDIS, quant à elles, n'évolueront plus. Si l'on souhaite des données plus récentes, il faut se tourner vers le successeur de SIDIS-greffe, SIDIS-suite, mais ce sont des données dont nous ne disposons pas aux fins de la présente recherche[11].

Tableau 1 – Vue générale des données

Caractéristique	**Données de condamnation**	**Données de détention**
Institution propriétaire des données	Service Public Fédéral Justice – Le Casier Judiciaire Central.	Service Public Fédéral Justice – La Direction Générale des Établissements Pénitentiaires (DG EPI).

[11] Il est toutefois prévu que les données de SIDIS-suite soient analysées dans un nouveau projet, lancé dans le prolongement du projet IIHA : le projet DOT (https://incc.fgov.be/DOT).

Nom du système informatique d'où les données ont été extraites	Casier judiciaire centra(a)l Strafregister (CJCS)	Système Informatique de Détention greffe (SIDIS-greffe) / Detentie Informatie Systeem griffie (SIDIS-griffie)
Acronyme utilisé dans ce rapport pour désigner le système information et la source de données correspondante	CJCS	SIDIS
Nombre de fichiers	64 fichiers	47 fichiers
Espace pris sur le disque par ces fichiers	5,6 GB	4,3 GB
Nombre approximatif de personnes concernées par ces données	Plus de 3 millions de personnes (concernées par plus de 8 millions de condamnations)	Plus de 300.000 personnes (concernées par plus de 700.000 détentions)
Date d'extraction des données	23 octobre 2020	10 octobre 2014
Période considérée dans les données[12]	1995-2020	1975-2014

2.4.1. Le Casier Judiciaire Central (CJCS)

Les tables qui composent l'extraction du Casier Judiciaire Central (CJCS) que nous avons reçue sont listées et résumées dans le Tableau 2 ci-dessous. Pour davantage d'information sur CJCS, il faut se rapporter à Huynen, Jeuniaux, et al. (2024).

Tableau 2 – Les tables du Casier Judiciaire Central (CJCS)

	Nom de la table	Nombre d'enregistrements	Signification de chaque enregistrement de la table
1	BULLETIN	8.271.193	Bulletin de condamnation d'une personne.

[12] Pour CJCS, bien qu'il y ait des enregistrements antérieurs à 1995, l'année 1995 est l'année à partir de laquelle on considère que les enregistrements sont complets.

2	DECISION	9.702.192	Décision prise dans le cadre d'un bulletin.
3	DOSSIER	3.860.989	Dossier d'une personne.
4	DOSSIER_MERGING	11.316	Trace historique de la fusion de deux dossiers.
5	DOSSIER_STAT	3.489.717	Nombre de tentatives de synchronisation avec le Registre National.
6	ECR_CENTRALAUTHORITY	28	Autorité centrale d'un pays.
7	ECR_CITY	2.287	Ville d'un pays.
8	ECR_COUNTRY	1	Un pays.
9	ECR_COUNTRYLANGUAGE	21	Langue d'un pays.
10	ECR_COUNTRYSUDIVISION	2.514	Région d'un pays.
11	ECR_CURRENCY	248	Devise d'un pays.
12	ECR_DECISIONCHANGETYPE	18	État actuel d'une décision.
13	ECR_LANGUAGE	21	Une langue.
14	ECR_OFFENCECATEGORY	188	Une catégorie d'infraction.
15	ECR_OFFENCELEVELCOMPLETION	3	Un niveau d'infraction.
16	ECR_OFFENCELEVELPARTICIPATIO	3	Participation dans un niveau d'infraction.
17	ECR_REQUESTPURPOSECATEGORY	40	Un type de demande.
18	ECR_SANCTIONALTERNATIVETYPE	3	Un type de sanction alternative.
19	ECR_SANCTIONCATEGORY	70	Une catégorie de sanction.
20	ECR_SANCTIONNATURE	3	Nature d'une sanction.
21	ECR_SANCTIONTYPESUSPENSION	5	Types de sursis.
22	FACT_CODE	11.194	Un code de fait.
23	FACT_CODE_MAP_ECRIS	572	Correspondance entre un code de fait CJCS et un code ECRIS.
24	FACT_CODE_RELATION	3.716	Relation entre deux codes de fait.
25	FACT_CODE_USE	16.816.738	Usage d'un code de fait pour décrire un fait.
26	FACT	13.383.607	Un fait.
27	HISTORY_RN_NR	4.999	Trace historique relative au numéro de registre national d'une personne.

28	IDENTITY	292.398	Identité d'une personne dans un dossier.
29	JURISDICTION	502	Une juridiction.
30	LEGAL_REMEDY	103.564	Trace historique relative à un recours.
31	LOV_APPLICATION_TYPE	3	Type de relation temporelle entre deux types de peines.
32	LOV_BULLETIN_STATUS	3	Statut d'un bulletin.
33	LOV_CITY	2.844	Une ville en Belgique.
34	LOV_CITY_TRANSLATION	19.643	Une ville européenne.
35	LOV_CIVIL_STATE	5	Un état civil.
36	LOV_COUNTRY	259	Un pays.
37	LOV_COUNTRY_TRANSLATION	768	Information sur un pays dans sa langue d'origine.
38	LOV_CURRENCY	248	Une devise.
39	LOV_DECISION_STATUS	2	Statut d'une décision.
40	LOV_DOSSIER_STATUS	4	Statut d'un dossier.
41	LOV_FACT_CODE_TYPE	3	Type de code de fait.
42	LOV_FACT_RELATION_TYPE	3	Type de relation entre deux faits.
43	LOV_IDENTITY_TYPE	2	Type d'identité (alias ou historique).
44	LOV_JUDGEMENT_DEGREE	3	Degré d'un jugement.
45	LOV_LEGAL_REMEDY_TYPE	13	Type de recours.
46	LOV_MUNICIPALITIES	607	Une commune belge.
47	LOV_NATIONALITY	207	Une nationalité.
48	LOV_NUMBER_OF_JUDGES	3	Un nombre de juges.
49	LOV_PUNISHMENT_CODE_TYPE	2	Type de code de peine ou mesure.
50	LOV_RN_STATUS	6	Statut du numéro de Registre national.
51	LOV_SERVICE	20	Un service ou tribunal.
52	LOV_SEX	3	Le sexe d'une personne.
53	NATIONALITY_DOSSIER	3.813.942	La nationalité associée à un dossier.
54	NATIONALITY_IDENTITY	101.491	La nationalité associée à une identité.

55	POSTAL_CODE	1.210	Un code postal.
56	PUNISHMENT_CODE	698	Un code de peine ou mesure.
57	PUNISHMENT_CODE_EXCLUSION	940	Une relation d'exclusion entre deux codes de peines ou mesures (codes qui ne peuvent pas être associés ensemble).
58	PUNISHMENT_CODE_MAP_ECRIS	225	Correspondance entre un code CJCS de peine ou mesure et un code ECRIS de peine ou mesure.
59	PUNISHMENT_CODE_RELATION	716	Relation entre deux codes de peines ou mesure.
60	PUNISHMENT_CODE_SPECIFIC	47	Une peine ou mesure spécifique.
61	PUNISHMENT	21.266.421	Une peine ou mesure.
62	PUNISHMENT_LIST_SPECIFIC	89.973	Lien entre une peine ou mesure d'un part, et une peine ou mesure spécifique.
63	PUNISHMENT_SUSPENSION	2.266.764	Information sur un sursis partiel.
64	PUNISHMENT_SUSPENSION_DATE	76.579	Information sur la durée d'un sursis.
65	RN_NATIONALITY	423	Information relative au traitement d'une mutation du numéro du Registre national.

2.4.2. SIDIS-greffe (SIDIS)

Les tables qui composent l'extraction de SIDIS-greffe (SIDIS) que nous avons reçue sont listées et résumées dans le Tableau 3 ci-dessous. Pour davantage d'information sur SIDIS, il faut se rapporter à Maes, Mine, et al. (2024).

Tableau 3 – Les tables de SIDIS-greffe (SIDIS)

	Nom de la table	**Nombre d'enregistrements**	**Signification de chaque enregistrement de la table**
1	AGENDA	17.178.321	Information relative à l'agenda d'un détenu par rapport aux procédures et à la validité des titres de détention (e.g., quand le détenu doit se présenter devant quel tribunal).
2	ALIASSEN	35.570	Lien entre deux numéros d'identification d'un même détenu.
3	BEHANDELD	0	Aperçu du membre du personnel qui a ajouté une certaine information.

4	CALCUL	328.456	Donnée utile pour le calcul de la peine.
5	CDES	2.700	Description d'un code (e.g., AB1 = homme, AB2 = femme).
6	CDTYPES	65	Description d'un type de code (e.g., AB = sexe).
7	CLASSIFICATION	114.635	Information visant à trouver un établissement pénitentiaire approprié pour le détenu et à préparer une éventuelle liste d'attente.
8	CONF_PERSO	126.546	Information relative à la conférence du personnel.
9	DELIT	1.773.959	Aperçu de la nature d'une infraction en vertu de laquelle la détention a été imposée.
10	DET_ANTERIEURE	100.253	Donnée sur une période de détention antérieure, prise en compte aux fins de détermination de la peine.
11	DETENTION	3.812.402	Information sur la détention.
12	DETGEVANG	3.642.592	Mouvement du détenu au cours d'une période de détention donnée.
13	FILTRE	137	Syntaxe des filtres utilisés dans l'application.
14	GESTION_PREVENTIVE	828.673	Information relative à la détention provisoire préventive.
15	GESTION_PROC	991	Information sur une procédure possible, chaque procédure ayant son propre mode de gestion.
16	GEVANGENISSEN	51	Un code prison.
17	GROUPE_CALCUL	32	Un groupe de titres de détention.
18	GROUPE_UTILISATEUR	9	Un groupe d'utilisateurs pouvant effectuer des contrôles (e.g., le gestionnaire du greffe ou l'employé du greffe désigné par un chef).
19	HIST_NATIONALITEIT	4.642	Trace historique relative au changement d'un code de nationalité (code EC). Par exemple, si le pays de naissance d'une personne est la Yougoslavie, il reste le pays de naissance même si ce pays n'existe plus.
20	INCIDENT	43.039	Information sur un incident survenu pendant une période de détention.
21	INTERRUPTION_PEINE	552.641	Information sur une interruption de peine.
22	JOUR_FERIE_OUVRE	76	Un jour férié (i.e., jour de fermeture supplémentaire).
23	JURIDICTION	1.145.605	Information relative à la juridiction et à l'instance judiciaire.
24	LIB_ANTICIPEE	24.818	Information sur la procédure de libération anticipée (VLV, VI, PV et ET).

25	MESURE	23.344	Information sur une mesure imposée.
26	MODALITE_PROP	184.119	Information sur la modalité de condamnation proposée.
27	MODALITE	20.791	Information sur le type d'arrestation.
28	MODELE_AGENDA	34	Ce qui peut être fait dans l'agenda.
29	MOTIF_RADIATION	858.623	Raison pour laquelle il a été mis fin à la détention.
30	OPMERKINGEN	2.036	Commentaire concernant le détenu.
31	PEINE	356.535	Information sur une peine prononcée.
32	PERMISSION_SORTIE	368.575	Information sur une autorisation de sortie.
33	POSTCODES	1.150	Un code postal
34	PROCEDURE_DIG	0	Information sur la procédure de service des cas individuels.
35	PROCEDURE	809.206	Information sur une procédure.
36	PROC_MINISTRE	292.066	Information sur une procédure pour le ministre.
37	REGIME	819.420	Information concernant un régime sous lequel une période de détention a été effectuée.
38	REGLE_GESTION	249	Une règle ou procédure de gestion.
39	REMARQUE	210.710	Commentaire concernant un détenu, associé à son titre de détention.
40	SIGNALETIEKEN	365.401	Fiche d'information relative à un détenu.
41	SITUATION_LEGALE	1.367.920	Situation légale (i.e., condition juridique) d'un détenu.
42	SORTIE	19.951	Information relative à une autorisation de sortie ou à un congé pénitentiaire.
43	STRAF_MAAT	51.386	Information sur la libération définitive en cause de mise en liberté provisoire.
44	TITRE_DETENTION	1.330.779	Un titre de détention.
45	TRADUCTION	673	Traduction en français et en néerlandais d'un terme spécifiquement utilisé dans les institutions pénitentiaires.
46	TYPE	48	Type de mouvement d'un détenu en détention.

47	VEILIGHEIDSCODES	187.748	Analyse des risques relatifs à une personne à transférer (i.e., transport, supervision du transport, etc.)

2.5. Les questions de recherche

Maintenant que le cadre de la recherche a été fixé, nous opérationnalisons les objectifs du travail en termes de questions de recherche.

Comme cela a été expliqué, un des grands objectifs de ce volet de la recherche est d'intégrer deux ensembles de données (CJCS d'une part et SIDIS d'autre part), en identifiant quels enregistrements de personnes dans ces données correspondent aux mêmes personnes.

Il y a 3.860.989 enregistrements de personnes dans CJCS. En théorie, ils correspondent tous à des personnes distinctes, mais dans la pratique, il y aura des enregistrements dupliqués, et par conséquent le nombre de personnes réelles dans CJCS sera inférieur à 3.860.989.

Il y a par ailleurs 365.401 enregistrements de personnes dans SIDIS. Pareillement que pour CJCS, en théorie, ces enregistrements correspondent tous à des personnes distinctes, mais dans la pratique, il y aura des enregistrements dupliqués, et par conséquent le nombre de personnes réelles dans SIDIS sera inférieur à 365.401.

Enfin, puisqu'il y aura un certain nombre de personnes qui ont à la fois un enregistrement dans SIDIS et dans CJCS, le nombre total de personnes sera inférieur à 3.860.989 + 365.401 = 4.226.390.

Combien trouvera-t-on de personnes dans les données ? Combien ont à la fois un enregistrement dans SIDIS et CJCS ?

Avant de pouvoir répondre à ce genre de questions, il faudra d'abord appliquer une procédure d'intégration des données afin reconnaître les personnes qui sont associées à ces différents enregistrements. La tâche d'intégrer les données doit surmonter différents obstacles.

Tout d'abord, nous ne disposons pas des « bonnes réponses » permettant de savoir à coup sûr quels enregistrements correspondent à quelles personnes. Il nous faudra uniquement utiliser l'information qui se trouve dans les données de CJCS et SIDIS, et accepter une part d'incertitude quant aux réponses que nous fournirons.

Ensuite, il sera difficilement réalisable de comparer tous les enregistrements les uns aux autres pour découvrir lesquels appartiennent aux mêmes personnes, à cause du très grand nombre des comparaisons possibles – il y en a en effet plus de 16 billions (voir Tableau 4). D'autres obstacles et considérations méthodologiques seront abordés dans la section 2.6.

Tableau 4 – Nombres de comparaisons possibles (A x B = C)

	Nombre de nœuds A	**Nombre de nœuds B**	**Nombre de comparaisons C**
intra-SIDIS	365.401	365.400	133.517.525.400
intra-CJCS	3.860.989	3.860.988	14.907.232.197.132
inter-SIDIS-CJCS	365.401	3.860.989	1.410.809.241.589
		TOTAL	16.451.558.964.121

En résumé, nous aurons deux types de question de recherche : des questions de nature méthodologique, et des questions sur les données. Dans le Tableau 5 ci-dessous, nous listons des exemples de questions de recherche de chaque type. Des questions plus précises seront posées à mesure que la matière est exposée. Les questions méthodologiques seront abordées dans la section 2.6 (« La procédure de création de l'IHD »). Les questions sur les données seront abordées dans la section 3 (« Résultats »).

Tableau 5 – Deux ensembles de questions de recherche

1	Questions sur la méthodologie	• Comment intégrer les données pour déterminer quels enregistrements sont associés à quelles personnes ? • Quelle stratégie mettre au point pour prendre en compte le grand nombre de comparaisons possibles ? • Comment gérer l'incertitude par rapport à l'attribution des enregistrements à telle ou telle personne ?
2	Questions sur les données	• Quelles sont les caractéristiques de ces personnes ? • Combien de personnes sont à la fois dans SIDIS et dans CJCS ?

Dans des publications ultérieures, nous nous chargerons de répondre à des questions se rapportant à la récidive et les carrières criminelles.

2.6. La procédure de création de l'IHD

La procédure de construction de l'IHD suit quatre grandes étapes, qui correspondent chacune à la construction d'éléments supplémentaires que l'on ajoute à la base de données. Ces étapes sont résumées dans le Tableau 6 ci-dessous. Après ce tableau, nous expliquons chaque étape en nous aidant d'illustrations.

Tableau 6 – Vue d'ensemble de la procédure de création de l'IHD

Étape	**Objectif de l'étape**	**Explication**
Étape 1	Construction des nœuds et relations de base	Sur la base d'une analyse des données sources (SIDIS et CJCS), les enregistrements des différentes tables sont convertis en nœuds et en relations. Parmi ces enregistrements on trouve des enregistrements sur les personnes, et notamment des enregistrements qui représentent en principe des personnes uniques (SIGNALETIEKEN dans SIDIS et DOSSIER dans CJCS). Nous appelons ces enregistrements les « enregistrements de personnes ».
Étape 2	Construction des nœuds d'intégration	On associe à chaque enregistrement de personne, des nouveaux nœuds contenant une version standardisée d'information personnelle pertinente liée à l'enregistrement de la personne (e.g., le prénom, le nom, la date de naissance). On appelle ces nouveaux nœuds les nœuds d'intégration. Ils sont reliés directement aux enregistrements de personnes.
Étape 3	Construction des liens d'intégration	Sur la base des nœuds d'intégration, on cherche des paires d'enregistrements de personnes candidats à la comparaison, parmi le très grand nombre de paires possibles. On examine en général ces paires de candidats en deux phases. Dans la première phase dite de dégrossissage, on examine les paires et on les évalue de manière sommaire pour voir si elles ont prometteuses. Si ce n'est pas le cas, on rejette la paire. Si c'est le cas, on passe à la deuxième phase, dite d'affinage, et on évalue les deux enregistrements de chaque paire d'une manière plus précise, là aussi sur la base des nœuds d'intégration. Si dans la deuxième phase il y a suffisamment de preuves pour faire une comparaison et qu'ils sont assez similaires, on établit un lien d'intégration entre eux.
Étape 4	Construction des nœuds de personnes	On va ici pouvoir utiliser les liens d'intégration construits à l'étape 3. On va choisir les enregistrements qui sont liés par des liens d'intégration, selon différents critères, de manière à identifier différents scénarios d'analyse. Sur la base de chaque scénario, on identifie quels enregistrements sont reliés entre eux mais isolés des autres enregistrements. Chaque sous-ensemble d'enregistrements interconnectés correspond à une personne, pour laquelle on crée un nouveau nœud « personne ».

2.6.1. Étape 1 : les nœuds et relations de base

À l'étape 1, les enregistrements des tables de SIDIS et CJCS ont été inspectés, et ont subi plusieurs analyses statistiques pour mieux comprendre la nature des données. Des éléments contextuels sur SIDIS et CJCS sont venus éclairer cette analyse tels que des descriptions des entités et relations de CJCS et SIDIS.

Pour pouvoir réaliser ces analyses, on a appliqué aux données sources des traitements préliminaires[13] visant à standardiser le format des fichiers et effectuer certaines corrections ou enrichissements sans impact sur la validité des données sources.

Sur la base de cette analyse, une modélisation en graphe particulière a été mise au point. La modélisation en graphe consiste à déterminer quels genres de nœuds et de relations – ainsi que les propriétés de ces nœuds et relations – seront créés en fonction des différentes tables des données de départ et les différents champs de ces tables. Dans la plupart des cas, un genre de nœud particulier correspond à une table particulière, et les types de relations suivent les indications du schéma entités-relations.

Par exemple, les genres de nœuds DOSSIER et BULLETIN correspondent aux tables DOSSIER et BULLETIN de CJCS, respectivement. Autrement dit, chaque enregistrement dans la table DOSSIER correspond à un nœud du genre DOSSIER dans le graphe, et chaque enregistrement dans la table BULLETIN correspond à un nœud du genre BULLETIN. Ensuite, puisqu'au sein de CJCS, chaque enregistrement de la table BULLETIN indique le numéro de dossier auquel il se rapporte (i.e., quel enregistrement dans la table DOSSIER), dans la modélisation en graphe, on va tout simplement tracer une relation orientée allant du nœud du genre BULLETIN au nœud du genre DOSSIER auquel il se rapporte.

La nature de cette analyse est décrite davantage dans la section 3.1 (voir Figure 13 et Figure 14).

Concrètement parlant, cet exercice de modélisation a conduit à transformer les données standardisées en fichiers d'importation afin de « populer » Neo4j (i.e., la remplir avec des données).

Une fois dans Neo4j, les données ont explicitement la forme d'un graphe, c'est-à-dire un ensemble composé de nœuds reliés entre eux par des relations. Les nœuds correspondent à des entités (i.e., à des objets particuliers, comme des personnes, des faits, des condamnations, des prisons, etc.). Les relations correspondent à des relations entre ces objets (e.g., une personne est liée à une prison dans la mesure où elle est emprisonnée dans cet établissement, une personne est liée à une peine dans la mesure où elle a été condamnée à cette peine).

Par ailleurs, les nœuds et les relations entre les nœuds sont caractérisés par des annotations qui permettent de récupérer et exploiter l'information dans la base de données. Les nœuds sont

[13] Par exemple, les entêtes de fichiers sont ajoutés quand ils sont absents, le même séparateur des colonnes (la tabulation) est utilisé pour tous les fichiers, et les enregistrements brisés (i.e., se retrouvant sur plusieurs lignes au lieu d'une) sont réparés (pour qu'ils figurent chacun sur une seule ligne).

caractérisés par des étiquettes (e.g., « personne », « fait », « condamnation », « prison »), tandis que les relations sont caractérisées par des types (e.g., « est emprisonné dans », « est condamné à »).

En guise d'exemples de nœuds créés dans Neo4j, nous représentons dans la Figure 1 ci-dessous des nœuds qui correspondent à des enregistrements des personnes. Certains enregistrements appartiennent à SIDIS et d'autres à CJCS. À l'étape finale (étape 4), il s'agira de décider quels enregistrements correspondent à quelles personnes.

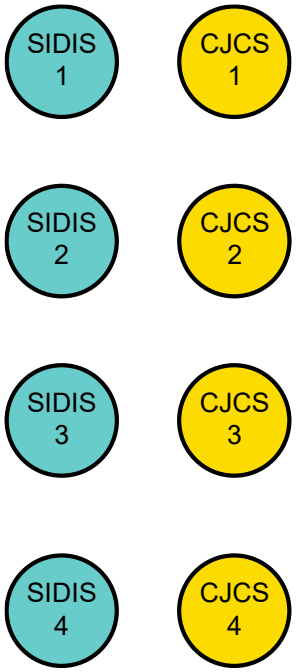


Figure 1 – Nœuds d'enregistrements de personnes

Dans la Figure 1, les nœuds correspondant à des enregistrements de personnes dans SIDIS sont ce qu'on appelle les fiches signalétiques des détenus (enregistrées originellement dans la table SIGNALETIEKEN de SIDIS), tandis que les nœuds correspondant à des enregistrements de personnes dans CJCS sont ce qu'on appelle des dossiers de condamnation (enregistrés originellement dans la table DOSSIER de CJCS).

2.6.2. Étape 2 : les nœuds d'intégration

À l'étape 2, on associe à chaque enregistrement de personne, des nouveaux nœuds contenant une version standardisée d'information personnelle pertinente liée à l'enregistrement de la personne (ex : le prénom, le nom, la date de naissance). On appelle ces nouveaux nœuds les nœuds d'intégration. Ils sont reliés directement aux enregistrements de personnes.

Les nœuds d'intégration

Pour illustrer cela, prenons un exemple fictif. Mettons que nous avons dans SIDIS l'enregistrement d'une personne britannique qui s'appelle James Bond et qui est née le 13 juin 1968. À l'étape 2, on crée les nœuds d'intégration correspondant (voir Figure 2).

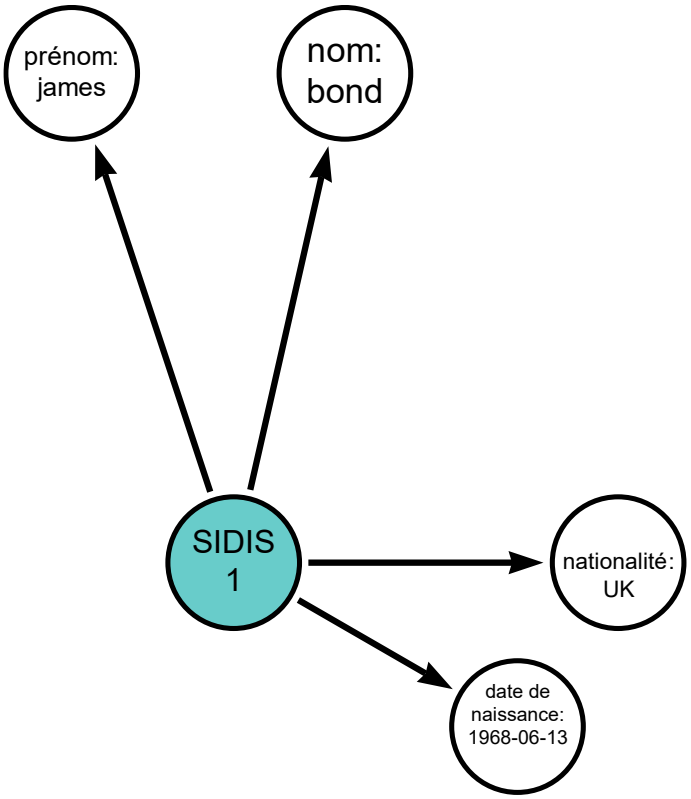


Figure 2 – Exemple fictif d'un enregistrement de SIDIS associé à des nœuds d'intégration

Mettons qu'il existe dans CJCS l'enregistrement d'une personne également britannique mais qui s'appelle Gems Bont, dont le numéro de registre national est « 007 » et dont la date de naissance est incomplète (voir Figure 3).

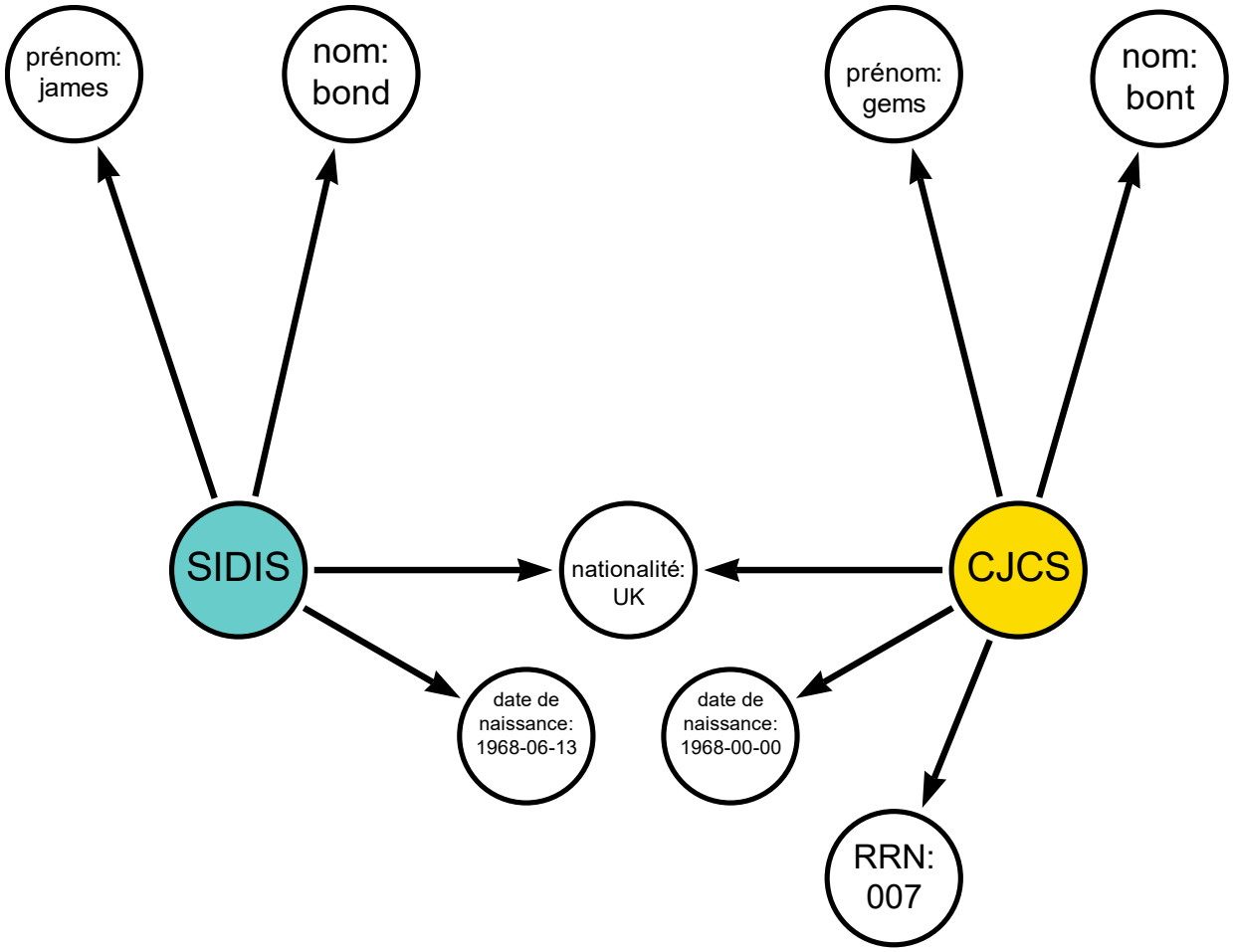


Figure 3 – Exemple fictif d'un enregistrement de SIDIS et CJCS associés à des nœuds d'intégration

S'agit-il de la même personne ? On voit que les deux enregistrements ont la nationalité en commun mais est-ce bien suffisant pour en décider ? Pour décider s'il s'agit de la même personne, nous ne sommes pas limités à l'information directement présente dans les enregistrements. En effet, on peut dériver de l'information supplémentaire à partir de l'information disponible. Par exemple, à partir de la date de naissance, on peut déterminer le jour, le mois et l'année de naissance. On peut aussi retenir les trois premières lettres du prénom et du nom, ainsi que les représentations phonétiques du prénom et du nom (voir Figure 4).

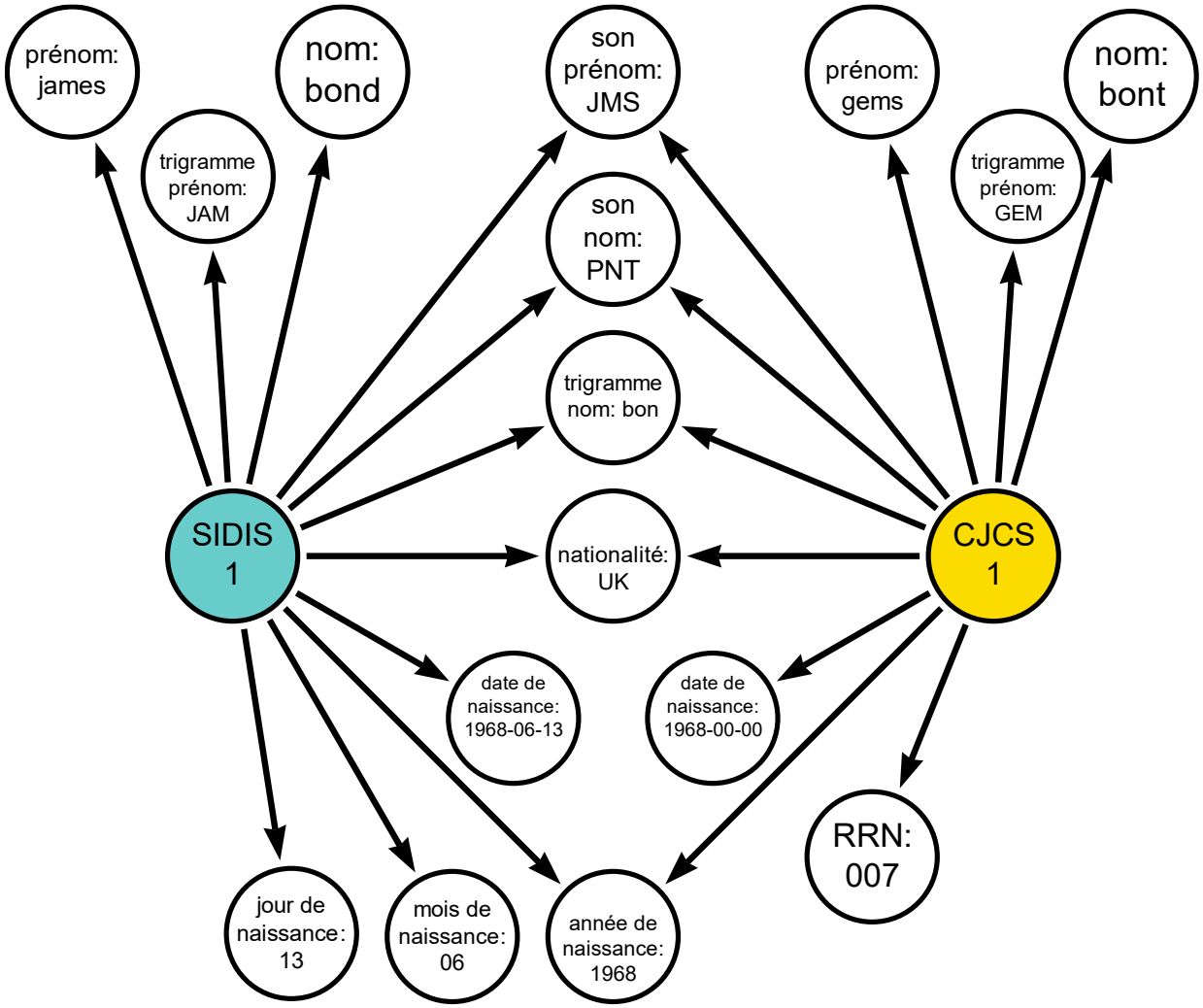


Figure 4 – Exemple fictif : ajout d'informations dérivées de la date de naissance, du prénom et du nom

Ces données supplémentaires sont des versions appauvries des données de départ, qui présentent l'avantage d'offrir davantage de points de comparaison au sein du graphe. En effet, si au départ (voir Figure 3), seul le nœud de la nationalité (UK) était en commun entre les deux enregistrements, on a à présent l'année de naissance (1968), ainsi que les trois premières lettres du nom de famille (bon), ainsi que la représentation phonétique du prénom (JMS) et du nom (PNT).

Vingt catégories de nœuds d'intégration

On a en tout 20 catégories de nœuds d'intégration, relatives à (1) le numéro de registre national (RRN), (2) le genre, (3) la nationalité, (4-6) le prénom, (7-9) le nom, (10-13) la date de naissance, (14-15) le pays et lieu de naissance, (16-19), la résidence, et (20) la date de jugement (voir Tableau 7).

Tableau 7 – Les nœuds d'intégration

	Nœud d'intégration	Explication
1	RRN	Le numéro de registre national. Deux enregistrements peuvent correspondre à la même personne et donc avoir le numéro de registre national en commun.
2	Genre	Le genre de la personne (homme, femme, non défini).
3	Nationalité	La nationalité de la personne, sachant qu'une personne peut avoir plusieurs nationalités.
4	Prénom	Le prénom de la personne. Une personne peut avoir plusieurs prénoms.
5	Trois premières lettres du prénom	Les trois premières lettres du prénom.
6	Représentation phonétique du prénom	La représentation phonétique du prénom via l'algorithme Double Metaphone.
7	Nom de famille	Le nom de famille de la personne. Une personne peut avoir plusieurs noms de famille si elle a plusieurs identités avec des noms de famille différents. Par ailleurs, les parties des noms de famille composés sont enregistrées dans des nœuds distincts (e.g., le nom de famille « Gold Smith » donnera lieu au nœud « Gold » et au nœud « Smith »).
8	Trois premières lettres du nom de famille	Les trois premières lettres du nom de famille.
9	Représentation phonétique du nom de famille	La représentation phonétique du nom via l'algorithme Double Metaphone.
10	Date de naissance	La date de naissance.
11	Année de naissance	L'année de naissance.
12	Mois de naissance	Le mois de naissance.
13	Jour de naissance	Le jour de naissance.

14	Pays de naissance	Le pays de naissance.
15	Lieu de naissance	Le lieu de naissance.
16	Pays de résidence	Le pays de résidence.
17	Lieu de résidence	Le lieu de résidence.
18	Code postal du lieu de résidence	Le code postal de résidence.
19	Adresse du lieu de résidence	L'adresse de résidence (e.g., un numéro et un nom de rue).
20	Date de jugement	La date de jugement au format (dd-MMM-yy, e.g., 18-JAN-02). Deux personnes peuvent avoir une ou plusieurs dates de jugement en commun.

Chaque nœud d'intégration résulte d'une extraction et standardisation des données de départ, et chaque nœud d'intégration est unique. Par conséquent, si deux enregistrements sont associés à un même nœud d'intégration, ces enregistrements sont liés indirectement l'un à l'autre du fait qu'ils partagent ce nœud. Par exemple, on a un nœud pour le mois de naissance en janvier dont la valeur est « 1 », et un nœud pour le mois de naissance en février dont la valeur est « 2 ». Si un enregistrement de personne dans SIDIS et un enregistrement de personne dans CJCS indiquent tous deux que la personne est née en janvier, leurs deux enregistrements seront indirectement reliés l'un à l'autre par le nœud du mois de naissance de janvier. C'est le même principe que celui de l'exemple avec la nationalité (voir Figure 3) et l'année de naissance (voir Figure 4).

La source des nœuds d'intégration

L'origine de l'information utilisée dans les bases de données sources (SIDIS et CJCS) pour créer ces nœuds d'intégration est indiquée dans le Tableau 8 ci-dessous. On voit par exemple que le nœud correspondant au numéro de Registre National (RRN) est trouvé dans le champ « rrn » de la table SIGNALETIEKEN de SIDIS, ainsi que dans les champs RN_NR des tables DOSSIER et HISTORY_RN_NR de CJCS.

Tableau 8 – Origine de l'information pour créer les nœuds d'intégration

	Nœud d'intégration	**SIDIS**	**CJCS**
1	RRN	Le champ **nrn** de la table SIGNALETIEKEN.	Les champs **RN_NR** des tables DOSSIER et HISTORY_RN_NR.

2	Genre	Le champ **cdab** de la table SIGNALETIEKEN.	Le champ **SEX_ID** de la table DOSSIER.
3	Nationalité	Le champ **cdec** de la table SIGNALETIEKEN.	Les champs **NATIONALITY_ID** des tables NATIONALITY_DOSSIER et NATIONALITY_IDENTITY.
4	Prénom	Le champ **voornamen** de la table SIGNALETIEKEN.	Les champs **FIRST_NAME_1**, **FIRST_NAME_2**, et **FIRST_NAME_3** des tables DOSSIER et IDENTITY.
5	Trois premières lettres du prénom	Le nœud **Prénom**.	Le nœud **Prénom**.
6	Représentation phonétique du prénom	Le nœud **Prénom**.	Le nœud **Prénom**.
7	Nom de famille	Le champ **naam** de la table SIGNALETIEKEN.	Les champs **SURNAME** des tables DOSSIER et IDENTITY.
8	Trois premières lettres du nom de famille	Le nœud **Nom de famille**.	Le nœud **Nom de famille**.
9	Représentation phonétique du nom de famille	Le nœud **Nom de famille**.	Le nœud **Nom de famille**.
10	Date de naissance	Le champ **gebdatum** de la table SIGNALETIEKEN.	Constituée sur la base des champs années, mois et jour de naissance ci-dessous.
11	Année de naissance	Le nœud **Date de naissance**.	Le champ **DATE_BIRTH_YEAR** de la table DOSSIER et le champ **BIRTH_DATE_YEAR** de la table IDENTITY.
12	Mois de naissance	Le nœud **Date de naissance**.	Le champ **DATE_BIRTH_MONTH** de la table DOSSIER et le champ **BIRTH_DATE_MONTH** de la table IDENTITY.

13	Jour de naissance	Le nœud **Date de naissance**.	Le champ **DATE_BIRTH_DAY** de la table DOSSIER et le champ **BIRTH_DATE_DAY** de la table IDENTITY.
14	Pays de naissance	Le champ **cdebgeboorte** de la table SIGNALETIEKEN.	Les champs **COUNTRY_OF_BIRTH_ID** des tables DOSSIER et IDENTITY.
15	Lieu de naissance	Le champ **gebplaats** de la table SIGNALETIEKEN.	Les champs **PLACE_BIRTH_ID** des tables DOSSIER et IDENTITY.
16	Pays de résidence	Le champ **cdeblandverblijf** de la table SIGNALETIEKEN.	Le champ **COUNTRY_ID** de la table DOSSIER.
17	Lieu de résidence	Le champ **gemeente** de la table SIGNALETIEKEN.	Les champs **POSTAL_CODE_ID** et **NIS_CODE_ID** de la table DOSSIER.
18	Code postal du lieu de résidence	Le champ **cdpost** de la table SIGNALETIEKEN.	Le champ **POSTAL_CODE_ID** de la table DOSSIER.
19	Adresse du lieu de résidence	Le champ **straat_nr** de la table SIGNALETIEKEN.	Les champs **ADDRESS** et **ADDITIONAL_ADDRESS** de la table DOSSIER.
20	Date de jugement	Le champ **date_jugement** de la table JURIDICTION.	Le champ **JUDGEMENT_DATE** de la table BULLETIN.

Opérations de standardisation

L'information trouvée dans les bases de données sources telle que rapportée dans le Tableau 8 n'est pas enregistrée à l'état brut en tant que nœud d'intégration : une certaine standardisation de l'information est effectuée afin de réduire la variabilité des valeurs et éliminer d'éventuelles anomalies. Ces opérations sont rapportées dans le Tableau 9 ci-dessous.

Tableau 9 – Méthode de standardisation pour créer les nœuds d'intégration

	Nœud d'intégration	**Séquences d'instructions suivies pour standardiser l'information et créer le nœud d'intégration**
1	RRN	(1) Suppression des caractères non alphanumériques ; (2) Taille (RRN) >= 8 caractères ; (3) Si taille (RRN) < 11 caractères, compléter le début du RRN avec des 0 ; (4) Le RRN est un entier > 0.
2	Genre	Dans le cas de SIDIS : Conversion du code AB utilisé pour suivre l'usage de CJCS (homme = M, femme = F, indéfini = U).
3	Nationalité	Dans le cas de SIDIS : Conversion des codes EC en code ISO représentant les pays, comme c'est le cas pour CJCS.
4	Prénom	Fonction CLEAN.TEXT[14] : Pour un mot donné : (1) Usage de la casse minuscule ; (2) Élimination des signes diacritiques et des éléments qui ne sont pas alpha numériques (i.e., dans [a-z] et [0-9]) ; (3) Renvoie le mot. Par exemple, « François » devient « francois » Fonction MOT.STANDARD : Pour un mot donné, (1) Applique CLEAN.TEXT (mais avec préservation de l'espace, du tiret et de l'apostrophe) ; (2) Retourne le mot, si Taille(mot) >1 et le mot n'est pas « inconnue », « onbekend » ou « zvn » (= zonder voornaam) ; (3) Renvoie le mot. Procédure : Pour le champ prénom (dans le cas de SIDIS) ou les trois champs prénom (dans le cas de CJCS) : (1) Élimination des espaces surnuméraires (e.g., doubles espaces) à l'intérieur et aux extrémités des chaînes ; (2) Via les espaces restants, découpage des mots distincts en prénoms distincts ; (3) Pour chaque prénom, applique MOT.STANDARD.

[14] https://neo4j.com/labs/apoc/4.1/overview/apoc.text/apoc.text.clean/

5	Trois premières lettres du prénom	Sur le prénom standardisé : Si le prénom est au moins de trois caractères, prendre les trois premiers caractères.
6	Représentation phonétique du prénom	Sur le prénom standardisé : Applique l'algorithme d'encodage phonétique double Metaphone de Neo4j[15]. Par exemple « joseph » devient « JSF ».
7	Nom de famille	Comme pour le prénom.
8	Trois premières lettres du nom de famille	Comme pour le prénom.
9	Représentation phonétique du nom de famille	Comme pour le prénom.
10	Date de naissance	Dans le cas de SIDIS : (1) Découper la date selon le format dd/mm/yyyy ; (2) Considérer 01/01/1901, et les dates dont le jour est en-dehors de [1-31] et le mois en-dehors de [1-12], comme invalides ; (3) Placer la date au format yyyy-mm-dd. Dans le cas de CJCS : (1) Assembler l'année, le mois et le jour de naissance pour former une date ; (2) Considérer comme invalide la date dont le jour est en-dehors de [1-31] et le mois en-dehors de [1-12] ; (3) Placer la date au format yyyy-mm-dd.
11	Année de naissance	Le traitement est similaire à celui de la date de naissance.
12	Mois de naissance	Le traitement est similaire à celui de la date de naissance.
13	Jour de naissance	Le traitement est similaire à celui de la date de naissance.
14	Pays de naissance	Dans le cas de SIDIS : Conversion des codes EB en code ISO représentant les pays, comme c'est le cas pour CJCS.
15	Lieu de naissance	Dans le cas de SIDIS :

[15] https://neo4j.com/labs/apoc/4.1/overview/apoc.text/apoc.text.doubleMetaphone/

		Pour un lieu donné : (1) Applique CLEAN.TEXT ; (2) Taille (lieu) >1 ; (3) Le lieu ne peut pas être une chaîne uniquement composée de X ou ?. Dans le cas de CJCS : Pour chaque nom de ville en français, néerlandais, anglais et allemand : Applique CLEAN.TEXT.
16	Pays de résidence	Comme pour le pays de naissance.
17	Lieu de résidence	Dans le cas de SIDIS : Pour un lieu donné : (1) Applique CLEAN.TEXT ; (2) Taille(lieu) >1 ; (3) Le lieu ne peut pas être une chaîne uniquement composée de « X » ou « ? ». Dans le cas de CJCS : Pour chaque nom de ville en français, néerlandais, anglais et allemand : Applique CLEAN.TEXT.
18	Code postal du lieu de résidence	Dans le cas de SIDIS : Pour un code donné : le code doit être un entier plus grand que 0. Dans le cas de CJCS : Pour un code donné : (1) Applique CLEAN.TEXT ; (2) le code doit être un entier plus grand que 0.
19	Adresse du lieu de résidence	Pour une adresse donnée : (1) Applique CLEAN.TEXT ; (2) Taille (adresse) > 1.
20	Date de jugement	Dans le cas de SIDIS : Conversion de la date en format dd/mm/yyyy (e.g., 18/01/2002) en format dd-MMM-yy (i.e., 18-JAN-02) de manière à avoir le même format utilisé que dans CJCS.

2.6.3. Étape 3 : les liens d'intégration

À l'étape 3, on exploite les nœuds d'intégration créé à l'étape 2 afin de décider si deux enregistrements de personnes concernent la même personne.

On cherche à créer non seulement des liens entre enregistrements de SIDIS et de CJCS (inter-SIDIS-CJCS), mais aussi des liens à l'intérieur de chaque source de données originale afin de détecter d'éventuels duplicatas. Ainsi, on crée d'une part des liens entre les enregistrements de SIDIS (intra-SIDIS), et d'autre part, des liens entre les enregistrements de CJCS (intra-CJCS).

Trouver des enregistrements candidats à comparer

Tout d'abord, il s'agit de trouver deux enregistrements de personnes candidats à la comparaison.

Pour trouver ces candidats à comparer on utilise les relations du graphe. Plus précisément, on souhaite examiner toutes les paires d'enregistrements de personnes qui partagent un ou plusieurs nœuds d'intégration en commun particuliers. Ces nœuds ont été préalablement sélectionnés pour leur capacité à trouver des paires prometteuses, c'est-à-dire parmi lesquelles se trouvent des enregistrements appartenant vraisemblablement à la même personne.

Par exemple, deux enregistrements pourront partager un nœud d'intégration se rapportant à une date de condamnation, ce qui signifient qu'ils se rapportent à des personnes qui ont été condamnées à la même date, e.g., le 14 avril 1988 (voir Figure 5).

Figure 5 – Exemple fictif : on s'apprête à comparer deux enregistrements parce qu'ils ont une date de condamnation en commun

Cette exploitation de nœuds en commun particuliers permet d'éviter de comparer toutes les paires d'enregistrements possibles (et ainsi éviter de réaliser les 16 billions de comparaisons possibles), en se concentrant sur des paires potentiellement utiles, et qui sont donc candidates pour la comparaison.

Six méthodes ont été conçues pour trouver des paires d'enregistrements candidats à la comparaison. L'utilisation de la date de condamnation appartient à l'une de ces six méthodes (voir Tableau 10). Pourquoi avoir conçu plusieurs méthodes pour trouver des enregistrements ? Parce que les enregistrements diffèrent l'un de l'autre. Ils n'ont en effet pas tous la même qualité et quantité de données. Certains ne sont pas associés à des dates de condamnation, n'ont pas de numéro de registre national en commun, etc.

La méthode M^{RRN} définit les candidats à comparer comme ceux qui ont au moins un numéro de registre national (RRN) en commun. Le RRN a l'avantage d'être très précis, puisqu'il s'agit d'identifier avec lui la personne de manière unique. Le nombre de candidats à comparer avec cette méthode est assez réduit. En effet, outre qu'il s'agit d'un numéro unique[16], comme nous le verrons dans la section sur les résultats, il n'est pas toujours disponible. Par ailleurs, il faut garder à l'esprit qu'il est possible que surviennent des erreurs d'attribution (numéro attribué à la mauvaise personne) ou des erreurs d'encodage (un mauvais numéro attribué à l'encodage).

[16] Il sera donc, par construction, rarement partagés entre deux enregistrements.

Tableau 10 – Six méthodes pour trouver des candidats à comparer

Méthode	**Éléments en commun**	**Explication**
M^{RRN}	RRN	A et B ont un numéro de Registre National en commun.
M^{TRI}	Trigrammes	A et B ont en commun les trois premières lettres d'un prénom et les trois premières lettres d'un nom.
M^{TRI-i}	Trigrammes inversés	A possède un prénom dont les trois premières lettres sont identiques aux trois premières lettres d'un nom de B, et A possède un nom dont les trois premières lettres d'un nom sont identiques aux trois premières lettres d'un prénom de B.
M^{PHO}	Représentations phonétiques	Comme pour les Trigrammes mais sur la base de la représentation phonétique du mot.
M^{PHO-i}	Représentations phonétiques inversées	Comme pour les Trigrammes inversés mais sur la base de la représentation phonétique du mot.
M^{JUG}	Date de jugement	A et B ont une date de jugement en commun.

La méthode M^{TRI} définit les candidats à comparer comme ceux qui ont au moins les trois premières lettres d'un nom et au moins les trois premières lettres d'un prénom en commun. Par « trigrammes » on entend donc ici les « trois premières lettres » du mot. La méthode M^{TRI} va générer un nombre de candidats bien plus élevé que M^{RRN}, car de nombreux enregistrements auront des trigrammes en commun. Elle présente l'avantage qu'elle est applicable, même lorsque le RRN n'est pas disponible. Mais encore faut-il que dans CJCS et SIDIS, les prénoms et noms d'une même personne commencent bien par les mêmes trois premières lettres. Si ce n'est pas le cas (par exemple, la personne s'appelle « Philippe » dans CJCS et « Filip » dans SIDIS), ces enregistrements ne figureront pas parmi les candidats à comparer.

Pour contourner le problème précédent, la méthode M^{PHO} est utilisée, qui utilise la représentation phonétique du mot. On utilise pour cela l'algorithme Double Metaphone[17]. Par exemple, les variantes « Philippe » et « Filip » sont toutes les deux représentées par « FLP ». La méthode M^{PHO} définit deux

[17] https://neo4j.com/labs/apoc/4.0/overview/apoc.text/apoc.text.doubleMetaphone/

candidats A et B tels que A possède un prénom ayant la même représentation phonétique qu'un prénom de B et possède un nom ayant la même représentation phonétique qu'un nom de B.

Ces méthodes présentent la faiblesse de ne pas prévoir le cas où un enregistrement a été encodé de manière erronée avec le prénom écrit à la place du nom et le nom à la place du prénom. S'il existe un autre enregistrement où cette erreur n'a pas été faite, il ne sera pas possible de relier les deux enregistrements, que ce soit via M^{TRI} ou M^{PHO}. Pour prévoir ce cas de figure, on utilise des versions « inversées »[18] de ces procédures où l'on considère que si l'on compare un enregistrement A à un enregistrement B, on va comparer le prénom de A au nom de B et le nom de A au prénom de B.

Une telle inversion donne lieu à M^{TRI-i} (trigrammes inversés) et M^{PHO-i} (représentations phonétiques inversées). La méthode M^{TRI-i} définit deux candidats A et B tels que A possède un prénom ayant les mêmes trois premières lettres qu'un nom de B et possède un nom ayant les mêmes trois premières lettres qu'un prénom de B. Pareillement, la M^{PHO-i} définit deux candidats A et B tels que A possède un prénom ayant la même représentation phonétique qu'un nom de B et possède un nom ayant la même représentation phonétique qu'un prénom de B. Aucun nouveau nœud d'intégration n'a été créé à l'étape 2 pour rendre cela possible. Ce sont les relations qui encodent l'inversion proprement dite.

Enfin, comme nous l'avons vu avec l'exemple de la Figure 5, la méthode M^{JUG} définit les candidats à comparer comme ceux qui ont au moins une date de jugement en commun. Puisque de nombreuses personnes pourront avoir été condamnées le même jour, une telle méthode va générer un très grand nombre de candidats possibles. Toutefois, elle ne sera valable que pour ces personnes dont la date de jugement a été encodée, or, comme nous le verrons en examinant les résultats, elle n'est pas toujours disponible.

On exploite donc ici les relations du graphe qui permettent d'identifier des éléments en commun. Il s'agit là d'une manière efficace d'identifier des candidats à comparer grâce aux propriétés de Neo4j. En effet, parce qu'elle stocke les relations de manière native en utilisant des pointeurs pour naviguer d'un nœud à l'autre, elle peut parcourir ces relations avec rapidité. Cependant, la nature même des données en commun qui permettent de réaliser ces rapprochements entre enregistrements, fait qu'un nombre extrêmement élevé de paires de candidats peut, selon les cas, être généré.

Or il faut encore pouvoir déterminer parmi ces paires de candidats lesquelles correspondent à des candidats qui sont suffisamment similaires l'un à l'autre. On a vu dans la Figure 5, qu'on a trouvé un enregistrement de SIDIS et un enregistrement de CJCS associé à la même date de condamnation du 14 avril 1988. Ces deux enregistrements concernent-ils la même personne ? Pour le déterminer, on les compare sur la base des nœuds d'intégration auxquels ils sont associés.

Une telle comparaison prendra un peu de temps. Afin de limiter le nombre de comparaisons effectuées, une stratégie opérant en deux phases a été mise au point : la phase 1 de dégrossissage, suivie de la phase 2 d'affinage (voir Tableau 11).

[18] Au sens commun de mettre les choses dans un sens contraire.

Tableau 11 – Stratégie de création de liens en deux phases

Phase	Nom de la phase	Ingrédients principaux
Phase 1	Dégrossissage	On utilise les propriétés du graphe pour trouver facilement des candidats. On compare les candidats via une mesure de similarité vectorielle. On ne retient que les meilleurs candidats (qui seront traités ensuite à l'étape suivante).
Phase 2	Affinage	On compare les candidats disponibles via une mesure de similarité textuelle. On ne retient que les meilleurs candidats. Ceux-ci permettront de définir des nœuds de personnes à l'étape 4.

Dans la phase 1, dite de dégrossissage, on utilise les nœuds d'intégration disponibles pour une comparaison sommaire mais rapide. Si cette comparaison indique qu'il y a suffisamment de similarité entre les deux enregistrements en termes de nœuds d'intégration, alors on passe à la phase 2, dite d'affinage, dans laquelle l'évaluation est plus sophistiquée mais plus lente. Cette manière de procéder en deux phases permet de gagner du temps en diminuant le nombre d'opérations à réaliser.

On notera, toutefois, que la phase 1 n'est pas appliquée pour la méthode M^{RRN}. En effet, étant donné le nombre réduit de liens créés par la méthode M^{RRN}, on passe immédiatement à la phase 2 du processus, sans passer par la phase 1. C'est-à-dire que dès que deux enregistrements possèdent un RRN en commun, on procède à leur évaluation via la procédure d'évaluation de la phase 2.

Phase 1 : dégrossissage

Dans la phase 1, dite de dégrossissage, on compare de manière grossière mais rapide les (éventuellement innombrables) candidats qui ont été ainsi rassemblés, et on ne retient que les paires de candidats les plus prometteuses. Pour toute paire d'enregistrements A et B à comparer, chaque enregistrement est représenté par un vecteur V contenant les valeurs des nœuds d'intégration.

L'ensemble des 20 catégories de nœuds d'intégration (voir Tableau 7) sont utilisées pour construire des vecteurs V de ce genre, à une exception près : la date de jugement n'est jamais utilisée[19]. Par ailleurs, si on utilise bien les catégories de nœuds se rapportant au nom et au prénom, on utilise les informations sur les prénoms et noms tels qu'ils sont encodés à l'origine dans les méthodes M^{JUG}, M^{TRI} et M^{PHO}, mais de manière « inversée » [20] pour les méthodes $M^{TRI\text{-}i}$ et $M^{PHO\text{-}i}$.

[19] La raison en est que certains enregistrements présentaient beaucoup trop d'informations de ce genre, ce qui occupait le vecteur au point de diminuer indument le niveau de similarité. Par conséquent deux enregistrements qui auraient dû être considérés comme identiques, passaient pour différents.

[20] Par exemple dans la méthode M^{RRN}, quand on compare deux enregistrements A et B, pour A on construit un vecteur où les noms proviennent des champs de noms et les prénoms proviennent des champs de prénoms.

Prenons un exemple où les vecteurs V^A et V^B contiennent les identifiants uniques des nœuds d'intégration *James*, *UK*, *homme*, et *Londres*.

$$V^A = [\,James, \quad UK, \quad homme\,]$$

$$V^B = [\,Londres, \quad UK, \quad homme\,]$$

La similarité des deux vecteurs V^A et V^B est ensuite évaluée via le coefficient de Jaccard[21]. Pour le calculer, on établit tout d'abord l'intersection des vecteurs V^A et V^B comme le vecteur contenant les valeurs communes aux deux vecteurs :

$$V^A \cap V^B = [\,UK, \quad homme\,]$$

On calcule ensuite l'union de V^A et V^B comme le vecteur contenant les valeurs qu'on trouve dans l'un ou l'autre de ces vecteurs :

$$V^A \cup V^B = [\,James, \quad UK, \quad homme, \quad Londres\,]$$

Le coefficient de Jaccard est le ratio entre la taille de l'intersection et la taille de l'union :

$$J(V^A, V^B) = \frac{|\,V^A \cap V^B\,|}{|\,V^A \cup V^B\,|} = \frac{2}{4} = 50\%$$

Il faut à présent expliquer comment utiliser une version pondérée du coefficient de Jaccard de façon à prendre en compte les situations où une valeur apparaît plusieurs fois dans un vecteur. De cette manière nous allons pouvoir donner plus de poids à certains nœuds d'intégration que d'autres dans le calcul du coefficient de similarité (voir Tableau 12).

Par exemple, si le poids est de 3 (comme pour le prénom), la valeur sera présente trois fois dans le vecteur V. Comme on le voit dans le Tableau 12, des informations que le RRN, le prénom, le nom, la date de naissance totalisent plus de poids par rapport aux autres éléments de preuve.

Tableau 12 – Poids des nœuds utilisés dans la phase de dégrossissage

	Nœuds d'intégration	**Poids[22]**
1	RRN	3
2	Genre	1
3	Nationalité	1
4	Prénom	3

Pareillement pour B. Tandis que dans la méthode $M^{TRI\text{-}i}$, quand on compare A et B, on ne procède pas ainsi. Pour A, on construit un vecteur où les noms proviennent des champs de noms et les prénoms des champs de prénoms. Mais pour B, c'est l'inverse : on construit un vecteur où les noms proviennent des champs de prénoms et les prénoms des champs de noms.

[21] Voir : https://neo4j.com/docs/graph-data-science/current/algorithms/similarity-functions/

[22] Quand le nœud d'intégration n'intervient pas, on indique « / » dans la ligne correspondante.

5	Trois premières lettres du prénom	/
6	Représentation phonétique du prénom	/
7	Nom de famille	3
8	Trois premières lettres du nom de famille	/
9	Repr. phonétique du nom de famille	/
10	Date de naissance	3
11	Année de naissance	1
12	Mois de naissance	1
13	Jour de naissance	1
14	Pays de naissance	1
15	Lieu de naissance	1
16	Pays de résidence	1
17	Lieu de résidence	1
18	Code postal du lieu de résidence	1
19	Adresse du lieu de résidence	1
20	Date de jugement	/

Comment utiliser dans le coefficient de Jaccard les occurrences multiples et quel est leur incidence ? Pour expliquer cela, nous prenons un autre exemple avec les vecteurs V^C et V^D, où cette fois le prénom *James* apparaît trois fois, en vertu du paramètre du Tableau 12.

$$V^C = [\,James, \quad James, \quad James, \quad UK, \quad homme\,]$$

$$V^D = [\,James, \quad James, \quad James, \quad UK, \quad Londres\,]$$

L'intersection de V^C et V^D contiendra cette fois-ci le minimum des occurrences de chaque identifiant qui apparaissent dans les deux vecteurs. Dans l'exemple, *James* apparaît minimum 3 fois, *UK* apparaît minimum 1 fois, tandis qu'*homme* et *Londres* apparaissent minimum 0 fois (et donc ils n'apparaissent pas dans l'intersection).

$$V^C \cap V^D = [\,James, \quad James, \quad James, \quad UK\,]$$

L'union de V^C et V^D contient le maximum des occurrences de chaque identifiant qui apparaissent dans les deux vecteurs. Ici *James* apparaît maximum 3 fois, tandis que *UK*, *homme* et *Londres* apparaissent chacun maximum 1 fois.

$$V^C \cup V^D = [\,James, \quad James, \quad James, \quad UK, \quad homme, \quad Londres\,]$$

La taille de l'intersection est de 4 et celle de l'union est de 6. Par conséquent, la similarité est de 4/6 (i.e., environ 67%).

$$J(V^C, V^D) = \frac{|\, V^C \cap V^D \,|}{|\, V^C \cup V^D \,|} = \frac{4}{6} \approx 67\%$$

Comme on peut le voir au travers de cet exemple, le fait de donner plus d'importance au nœud d'intégration *James* (qui apparaît trois fois au lieu d'une), conduit à une augmentation du taux de similarité de 50% à 67%. En effet, le numérateur augmente plus (il passe de 2 à 4 : une augmentation de 100%) que le dénominateur (qui passe de 4 à 6 : une augmentation de 50%).

À contrario, cela conduira à une diminution du taux de similarité si l'un des vecteurs possède une valeur répétée qui ne se retrouve pas dans l'autre vecteur. Nous en donnons un exemple ci-dessous où James est répété trois fois dans V^C mais pas dans V^E.

$$V^c = [\, James, \quad James, \quad James, \quad UK, \quad homme \,]$$

$$V^E = [\, UK, \quad Londres \,]$$

$$V^C \cap V^E = [\, UK \,]$$

$$V^C \cup V^E = [\, James, \quad James, \quad James, \quad UK, \quad homme, \quad Londres \,]$$

$$J(V^C, V^E) = \frac{|\, V^C \cap V^E \,|}{|\, V^C \cup V^E \,|} = \frac{1}{6} \approx 17\,\%$$

On voit dans cet exemple que le dénominateur ne change pas (il reste à 6), tandis que le numérateur diminue (il passe de 4 à 1). On obtient ainsi une similarité de 1/6 (i.e., environ 17 %).

Le coefficient de Jaccard est relativement sommaire dans la mesure où il ne calcule pas la similarité des valeurs des vecteurs proprement dites (i.e., on ne compare pas *James* à *UK*, *James* à *homme*, etc. pour vérifier s'ils sont similaires ou pas), mais bien la similarité des vecteurs entre eux.

Quoi qu'il en soit, la mesure de similarité qui en résulte sera un nombre situé entre 0% et 100%. Si la similarité atteint un certain seuil, les deux enregistrements sont jugés comme suffisamment similaires et seront examinés dans la phase 2 d'affinage. Ce seuil a été fixé au seuil de 30%.

Dans l'exemple précédent, puisque $J(V^A, V^B) = 50\%$, $J(V^C, V^D) \approx 67\%$, et $J(V^C, V^E) \approx 17\%$, seules les paires (V^A, V^B) et (V^C, V^D), qui passent le seuil de 30%, seront examinée à la phase 2.

Phase 2 : affinage

Dans la phase 2, dite d'affinage, on examine parmi les paires de candidats disponibles, lesquelles sont les plus prometteuses en les soumettant à une analyse plus fine que celle qui a eu lieu à la phase 1 de dégrossissage. Elle est plus fine, mais elle est aussi plus lente, ce qui justifie l'existence de la phase 1 pour s'attaquer à un nombre (éventuellement excessif) de comparaisons. Comme nous l'avons déjà expliqué, cette stratégie a été appliquée pour toutes les méthodes sauf la M^{RRN}, étant donné le nombre réduit de candidats à comparer qui sont produits grâce au RRN. Pour M^{RRN}, on applique directement la phase « d'affinage » sur tous les enregistrements qui présentent un RRN en commun.

Dans la phase 2 d'affinage, pour toute paire d'enregistrement A et B à comparer, A est comparé à B sur chaque catégorie de nœud qu'il a en commun (date de jugement, RRN, etc.). Par exemple, si A possède

un prénom et B possède un prénom[23], on vérifie s'ils sont identiques. S'ils ne le sont pas, on vérifie s'ils sont semblables via la mesure de similarité textuelle de Jaro-Winkler[24]. S'ils n'atteignent pas sur cette mesure le seuil de similarité fixé (voir Tableau 13), on dit qu'ils sont différents.

Tableau 13 – Paramètres utilisés dans la phase d'affinage

	Nœuds d'intégration[25]	**Poids si identique**	**Seuil de similarité**	**Poids si similaire**
1	RRN	6	100%	0
2	Genre	2	100%	0
3	Nationalité	2	100%	0
4	Prénom	4	90%	2
5	Trois premières lettres du prénom	/	/	/
6	Repr. phonétique du prénom	/	/	/
7	Nom de famille	6	90%	3
8	Trois premières lettres du nom	/	/	/
9	Repr. phonétique du nom de famille	/	/	/
10	Date de naissance	4	100%	0
11	Année de naissance	3	100%	0
12	Mois de naissance	2	100%	0
13	Jour de naissance	2	100%	0
14	Pays de naissance	2	100%	0
15	Lieu de naissance	2	80%	1
16	Pays de résidence	1	100%	0
17	Lieu de résidence	1	80%	0,5
18	Code postal du lieu de résidence	1	100%	0
19	Adresse du lieu de résidence	1	70%	0,5
20	Date de jugement	1	100%	0

[23] Dans le cas où l'on a plusieurs prénoms, par exemple deux prénoms pour A et trois prénoms pour B, on effectue toutes les comparaisons entre les deux prénoms de A et les trois prénoms de B, et on garde le meilleur résultat du point de vue de l'enregistrement qui possède le moins de valeurs (dans cet exemple, l'enregistrement A).

[24] https://pypi.org/project/jaro-winkler/

[25] Quand le nœud d'intégration n'intervient pas, on indique « / » dans la ligne correspondante.

Cette procédure est effectuée pour chaque catégorie de nœud d'intégration commun à A et B (date de jugement, RRN, etc.), et ce dans le but de construire une somme de poids S, en sommant les poids obtenus pour chaque catégorie. Si pour une catégorie particulière, on trouve deux valeurs identiques (e.g., deux dates de jugement identiques ou deux RRNs identiques), on ajoute à la somme S le poids qui correspond au constat qu'elles sont identiques ; s'il y a deux valeurs semblables, on ajoute à la somme S le poids qui correspond au constat qu'elles sont similaires (voir Tableau 13).

Si pour toutes les comparaisons effectuées, on avait observé des valeurs identiques, on aurait eu une somme de similarité maximale MAX(S). La valeur de MAX(S) se situe entre zéro et quarante[26]. Dans la réalité, on a généralement $S < MAX(S)$. Le ratio des deux valeurs définit la similarité normalisée $SIM = \frac{S}{MAX(S)}$ qui se situe entre 0% et 100%.

À la fin de la comparaison de A et B, on observe la similarité normalisée SIM et la similarité maximale MAX(S). La similarité normalisée SIM représente le taux de similarité des enregistrements. Plus il est proche de 100%, et plus ils sont similaires. La similarité maximale représente le poids des preuves qui ont été examinées pour aboutir à cette mesure de similarité. Plus il est proche de 40, et plus on a été en mesure d'examiner des preuves pour juger de la similarité des enregistrements.

Par simplicité d'expression, dans le reste du rapport, on se réfère à MAX(S) comme étant le « poids des preuves », ou plus simplement « le poids ». Enfin, quand nous parlons simplement de la « similarité » des enregistrements, nous faisons référence à la mesure de similarité normalisée SIM.

Dans le cas de la M^{RRN}, on trace un lien d'intégration entre les enregistrements A et B et on note sa similarité SIM et son poids MAX(S). Dans le cas des autres méthodes (M^{TRI}, M^{PHO}, etc.), étant donné le grand nombre de paires de candidats comparées, on n'établira un lien d'intégration entre A et B que si SIM a atteint le seuil de 50% et MAX(S) le seuil de 10. Ces valeurs ont été choisies sur une base pragmatique, après plusieurs essais et erreurs, en espérant capturer un maximum de liens de bonne qualité, tout en limitant le nombre de liens d'intégration créés, sachant que descendre ces seuils reviendrait à augmenter le nombre de liens de manière excessive[27].

Si donc la comparaison réalisée en phase 2 est satisfaisante (au sens de satisfaire le seuil de SIM à 50% et MAX(S) à 10), un lien d'intégration est tracé.

Pour conclure

Prenons un exemple pour caractériser l'ensemble de l'étape 3.

---

[26] La valeur de MAX(S) est au plus 40. Le nombre 40 représente la somme de tous les poids dans le cas d'un constat d'identité des valeurs (voir Tableau 13). Dans la réalité, on n'observe pas toujours ce poids de 40 car A et B n'ont pas toujours des éléments d'une même catégorie en commun à comparer. Par exemple, si A possède un RRN, mais B n'a pas de RRN, on ne peut faire de comparaison sur ce type de nœud d'intégration et le poids correspondant n'est pas intégré au dénominateur de la similarité normalisée SIM.

[27] Plus on crée des liens et plus cela prend de l'espace sur le disque et plus cela prend du temps pour les créer et ensuite les effacer dans le cas où l'on n'en a plus besoin.

Imaginons que l'on applique la méthode M^{JUG} pour tracer des liens inter-SIDIS-CJCS. En parcourant le graphe on a trouvé deux enregistrements à comparer, qui, comme dans la Figure 5, partagent la même date de condamnation : le 14 avril 1988. Il s'agit d'un certain James Bond de SIDIS et un certain Gems Bont de CJCS, au sujet desquels on a également d'autres informations connues grâce aux autres nœuds d'intégration (date de naissance, nationalité, etc.).

Dans l'étape 1 de dégrossissage, on constitue des vecteurs avec ces informations en prenant en compte leur poids (voir Tableau 12). On compare ces deux vecteurs via le coefficient de Jaccard, et on obtient 40%, ce qui est supérieur au seuil de 30%. Par conséquent on passe à l'étape 2 d'affinage.

Dans l'étape 2 d'affinage, on utilise encore une fois les nœuds d'intégration pour comparer les deux enregistrements mais cette fois en utilisant la mesure de similarité textuelle de Jaro-Winkler, afin de construire un score de similarité en additionnant des poids pour chaque type de nœud d'intégration (voir Tableau 13). Mettons qu'on avait pour chaque enregistrement, tous les types de nœuds d'intégration, sauf le RRN qui vaut 6. On a donc MAX(S) = 34 (et non 40). Mettons qu'en comparant ces nœuds, on obtient un score S de 22. La similarité normalisée est donc SIM = 22 / 34 = 65%.

Puisque MAX(S) est supérieur à 10 et SIM est supérieur à 50%, on trace un lien d'intégration (voir Figure 6).

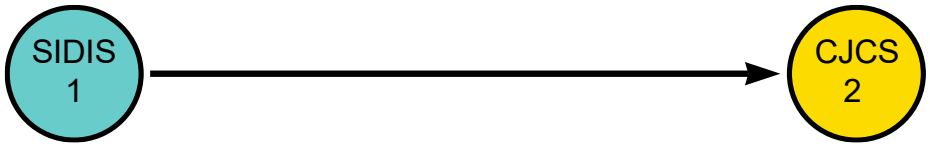


Figure 6 – Exemple fictif : création d'un lien d'intégration entre deux enregistrements

En général, puisque six méthodes (M^{RRN}, M^{TRI}, M^{PHO}, etc.) ont été utilisées pour trouver des candidats et les comparer, jusqu'à six liens d'intégration peuvent être tracés entre deux enregistrements (voir Figure 7). L'illustration de la Figure 7 sous-entend qu'on a tracé six liens d'intégration entre eux.

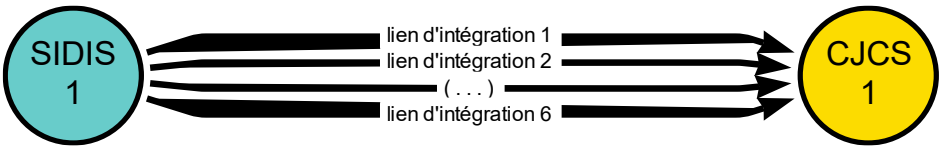


Figure 7 – Exemple fictif : création de liens d'intégration selon les six méthodes envisagées

À la fin de l'étape 3, nos nœuds d'enregistrements de personnes seront donc reliés ou non par des liens d'intégration. Pour poursuivre l'illustration commencée avec la Figure 1 au tout début, nous avons ajouté des liens d'intégration dans la Figure 8 ci-dessous. Par simplicité, nous nous sommes bornés à n'indiquer qu'un seul lien d'intégration par paire de nœuds et non pas six comme dans la Figure 7.

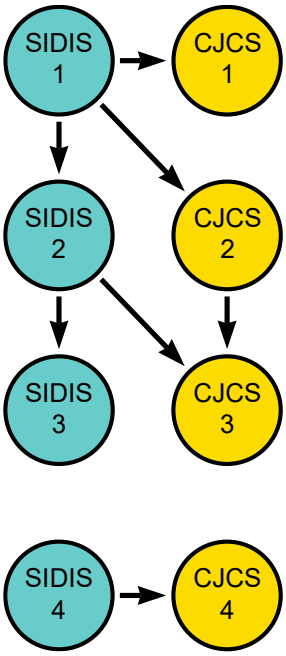


Figure 8 – Nœuds d'enregistrements de personnes reliés par des liens d'intégration

On voit dans la Figure 8 que l'enregistrement SIDIS 1 est lié à trois enregistrements (SIDIS 2, CJCS 1, CJCS 2). Par ailleurs SIDIS 2 est lié aux enregistrements SIDIS 3 et CJCS 3, et CJCS 2 est également lié à CJCS 3. Il est donc possible que ces six enregistrements appartiennent à la même personne. Enfin, SIDIS 4 est lié à CJCS 4, et donc il est possible que ces deux derniers enregistrements appartiennent à la même personne, mais une personne différente que la personne précédente. Comme on le voit dans la figure, on a deux sous-ensembles distincts de nœuds connectés dans le graphe. Le premier sous-ensemble de nœuds connectés l'un à l'autre comprend les nœuds de SIDIS 1 à SIDIS 3 et de CJCS 1 à CJCS 3. Le second sous-ensemble de nœuds connectés l'un à l'autre comprend les nœuds SIDIS 4 et CJCS 4.

Toutefois, ce n'est pas parce qu'il y a un lien entre deux enregistrements que ce lien sera *in fine* retenu au terme de la procédure. Un tel examen des liens sera effectué à l'étape 4.

2.6.4. Étape 4 : les nœuds de personne

À l'étape 4, on examine les liens d'intégration créés à l'étape 3 et on applique différents critères afin d'établir différents scénarios d'analyse. Par exemple, on sélectionnera les liens d'intégration qui ont atteint un certain seuil de similarité et de poids, et on écartera ceux qui n'ont pas atteint ce seuil. En reprenant l'exemple de la Figure 8, on marque en rouge dans la Figure 9 les liens qui seront écartés car ils n'atteignent pas le seuil retenu.

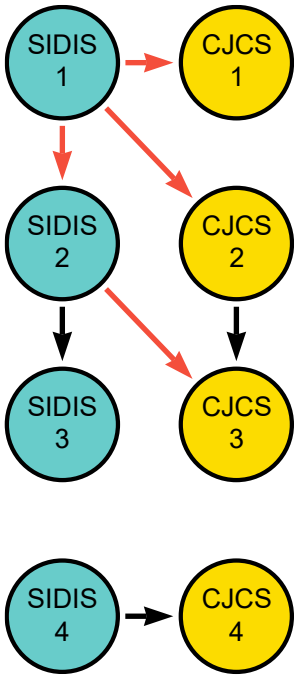


Figure 9 – Identification des liens jugés trop faibles (en rouge dans le graphique)

La Figure 10 ci-dessous correspond à la Figure 9 après avoir retiré les liens considérés comme trop faibles.

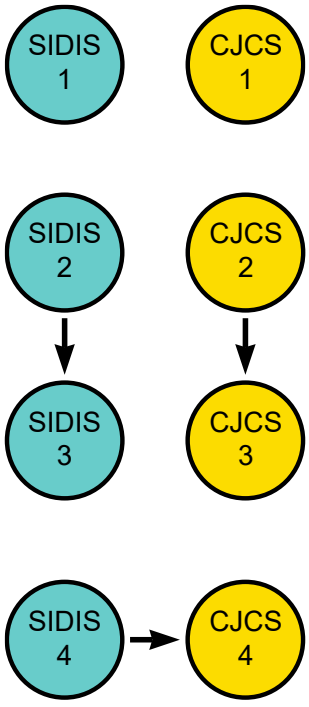


Figure 10 – Nœuds d'enregistrements de personnes reliés par des liens d'intégration (après avoir ignoré des liens jugés trop faibles)

Pour déterminer qu'on a affaire à des personnes distinctes, on va trouver les sous-ensembles de nœuds connectés entre eux. Des nœuds forment un « sous-ensemble » [28] quand on peut « passer » de l'un à l'autre en « traversant » un lien, mais qu'on ne peut pas « atteindre » d'autres nœuds de la base de données via un lien. On « passe » d'un nœud à l'autre, en « traversant » un lien, pour « atteindre » un nœud : le vocabulaire employé ici fait appel à la métaphore d'une carte routière via laquelle on cherche son chemin, pour passer d'une ville (nœud de départ) à une autre (un autre nœud) en passant par des routes (des liens), afin d'atteindre la ville souhaitée (nœud d'arrivée). Les individus sont comme les îles d'un archipel. Chaque individu correspond à une île sur laquelle les différents villages sont reliés l'un à l'autre par des routes. Mais il n'est pas possible d'atteindre les villages des autres îles de l'archipel, car il n'y a pas de route sur la terre ferme pour passer d'une île à l'autre. L'idée centrale est de trouver chaque île dans l'archipel : trouver chaque sous-ensemble formé par un ou plusieurs enregistrements interconnectés censés appartenir à un seul et même individu.

Examinons quelques exemples via la Figure 10. Par suite de l'élimination de liens jugés trop faibles (voir Figure 9), seuls trois liens apparaissent comme valides : un lien entre SIDIS 2 et SIDIS 3, un lien entre CJCS 2 et CJCS 3 et un lien entre SIDIS 4 et CJCS 4. À partir de SIDIS 2 on peut atteindre SIDIS 3 mais aucun autre nœud dans le graphe : SIDIS 2 et SIDIS 3 forment donc un sous-ensemble. SIDIS 1 n'est lié à aucun autre nœud : il forme donc un sous-ensemble à lui tout seul. Il en est de même pour CJCS 1. Au total, dans la Figure 10, on a cinq sous-ensembles de nœuds connectés : SIDIS 1 et CJCS 1 sont isolés (ce qui donne deux sous-ensembles), tandis que SIDIS 2 est associé à SIDIS 3, CJCS 2 est associé à CJCS 2 et SIDIS 4 est associé à CJCS 4, ce qui donne trois sous-ensembles de plus.

Puisque chaque sous-ensemble de nœuds est censé appartenir à un seul et même individu, aux cinq sous-ensembles distingués dans la Figure 10 correspondent cinq nœuds de personnes distincts. On ajoute donc dans la figure les cinq nœuds de personnes, chaque nœud de personne pointant vers son ou ses enregistrements (voir Figure 11).

[28] En théorie des graphes, on appellera ces sous-ensembles des « composants » ('components' en anglais) ou « composants connexes ».

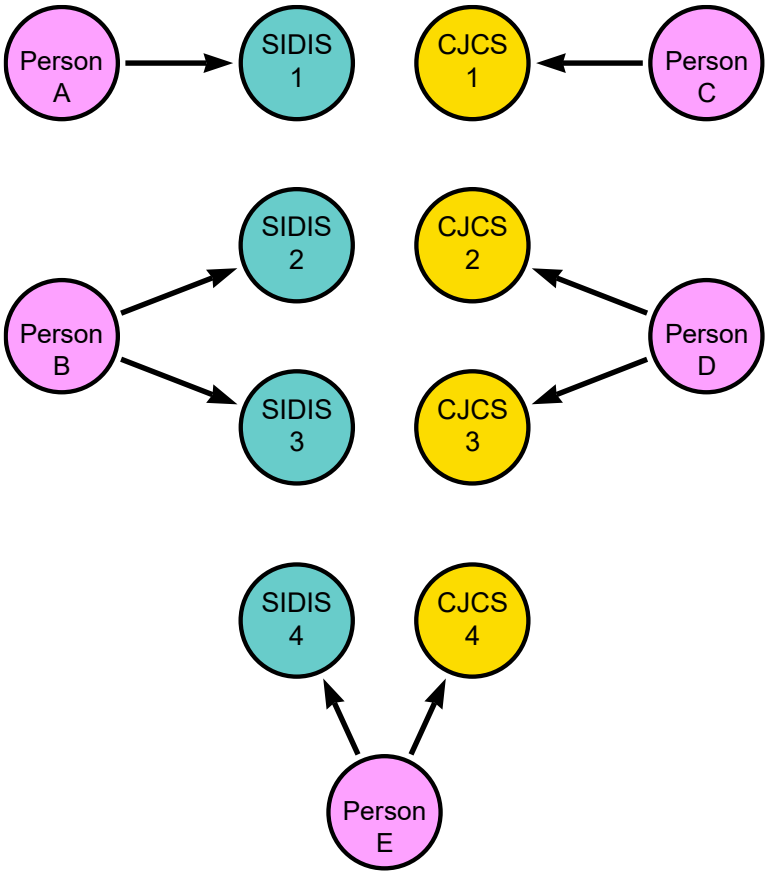


Figure 11 – Nœuds d'enregistrements de personnes liés à des nœuds de personnes

On observe qu'un nœud de personne est créé, même pour un enregistrement isolé. En effet une personne A est associée à l'enregistrement SIDIS 1 et une personne C est associée à l'enregistrement CJCS 1, qui sont tous les deux des enregistrements isolés.

La personne B présente un enregistrement en double (deux enregistrements dans SIDIS), de même que la personne D (deux enregistrements dans CJCS).

Enfin la personne E associée aux enregistrements SIDIS 4 et CJCS 4. C'est ce cas de figure qui permet d'étudier la récidive d'une personne conjointement avec les données de condamnation (de CJCS) et les données de détention (de SIDIS).

Si l'on devait procéder à l'identification des sous-ensembles pour la Figure 8, qui contient plus de liens, on n'aurait pas cinq sous-ensembles de nœuds mais uniquement deux sous-ensembles, et donc uniquement deux nœuds de personnes.

Comme l'illustre cet exemple fictif, la nature des sous-ensembles d'enregistrements correspondant à une personne et donc le nombre de nœuds de personnes ainsi créés, dépendront étroitement des choix méthodologiques effectués.

À l'étape 4, on va donc créer des nœuds de personnes sur la base des liens créés à l'étape 3 entre les 4.226.390 enregistrements de personnes. La manière dont on va choisir ces enregistrements de personnes et les liens qui les unissent, va fortement influencer les nœuds de personnes créés. Nous avons effectué différents choix méthodologiques en ce sens, que nous avons exprimé au travers de

l'application de six critères (voir Tableau 14). Ces critères sont brièvement énoncés dans le tableau et discutés ci-après.

Tableau 14 – Critères utilisés pour retenir les nœuds et liens afin de créer des nœuds de personne

Critère	**Explication**	**Nombre de conditions**
Critère 1	Peu importe quelle méthode a été utilisée pour tracer un lien d'intégration (M^{RRN}, M^{TRI}, etc.). Dès qu'un lien d'intégration existe, il peut être exploité.	1
Critère 2	On utilise toujours les liens établis dans CJCS entre dossiers et dossiers fusionnés.	1
Critère 3	On ne cherche pas explicitement à exclure les personnes morales de CJCS.	1
Critère 4	Six combinaisons de seuils de similarité et de poids de preuves.	6
Critère 5	Dans CJCS, uniquement les dossiers actifs ou aussi les inactifs ?	2
Critère 6	Uniquement liens inter-SIDIS-CJCS ou aussi intra-SIDIS et intra-CJCS ?	2

Comme nous le verrons au terme de cette explication, le croisement de ces six critères a mené à la définition de 24 scénarios méthodologiques différents. L'idée de scénario méthodologique renvoie bien au fait que des choix méthodologiques ont été posés, et qu'ils sont susceptibles d'influencer les résultats. Il s'agit donc d'un appel à la prudence lorsqu'on examine ces résultats.

Enfin, bien qu'une multiplicité de scénarios complexifie l'analyse, l'intérêt d'en envisager plusieurs est de pouvoir conduire, pour chacun d'eux, les analyses statistiques qui nous intéressent. Nous sommes alors en mesure de montrer comment les résultats de ces analyses varient (ou pas) en fonction des scénarios envisagés. Moins le résultat variera et plus nous aurons confiance dans l'estimation ainsi produite, en faisant la supposition que parmi les scénarios considérés se trouve un scénario suffisamment acceptable pour les usages escomptés. Mais si le résultat varie, il nous faudra comprendre pourquoi c'est le cas et prendre les mesures qui s'imposent.

Le critère 1

Le critère 1 concerne la nature du lien d'intégration. À l'étape 3, six méthodes ont été appliquées pour tracer des liens d'intégration : M^{RRN}, M^{JUG}, M^{TRI}, M^{PHO}, $M^{TRI\text{-}i}$, $M^{PHO\text{-}i}$ (voir Tableau 10). À l'étape 4, sauf mention contraire, on ne fait pas de distinction entre ces six méthodes dans le sens où dès qu'un lien a été tracé, il est susceptible d'être utilisé, et peu importe laquelle des six méthodes a permis de tracer

ce lien. Ce critère est appliqué par défaut. Il s'agit donc d'une constante à travers l'ensemble des 24 scénarios considérés.

Le critère 2

Le critère 2 concerne les liens établis dans CJCS entre les dossiers et les dossiers fusionnés. Nous allons toujours considérer un dossier de CJCS et ses dossiers « fusionnés » comme étant liés l'un à l'autre. Ces dossiers « fusionnés » sont des dossiers inactifs dont l'information a été transposée aux dossiers actifs (voir Huynen, Jeuniaux, et al., 2024). En d'autres mots, qu'il y ait ou non un lien d'intégration créé par l'étape 3 entre les deux enregistrements, du fait qu'il y a un lien logique entre eux (tel qu'encodé dans CJCS), ils feront partie du même sous-ensemble de nœuds. Cela signifie qu'un dossier A et son dossier fusionné B feront toujours partie du même sous-ensemble et qu'ils ne pourront pas se retrouver dans des sous-ensembles séparés. Ce critère est toujours appliqué. Il s'agit donc là aussi d'une constante à travers l'ensemble des 24 scénarios considérés.

Le critère 3

Le critère 3 concerne la nature des dossiers de personnes. Si CJCS contient en grande majorité des dossiers de personnes physiques (c'est-à-dire des personnes au sens commun du terme), elle contient également des dossiers de personnes morales, c'est-à-dire des entreprises ou associations (environ 43.000). Dans l'algorithme d'intégration, aucune attention particulière n'est apportée à la distinction entre les dossiers de personnes physiques et les personnes morales. L'analyse de cet aspect n'est réalisée qu'a posteriori, une fois que les nœuds de personnes ont été créés. Par conséquent, par nœud de « personne » nous entendons aussi bien une personne physique que morale. Il est toutefois certain qu'aucune personne morale ne pourra se retrouver dans SIDIS car les personnes morales ne font naturellement pas de la détention. On s'attend donc à ce que notre procédure d'intégration n'ait pas mené à lier un enregistrement de CJCS correspondant à une personne morale à un enregistrement de SIDIS correspondant à une personne (nécessairement) physique. Ce sera le cas dans la grande majorité des cas, bien qu'aucun mécanisme dans l'algorithme n'empêche de commettre une telle erreur. Ce critère est toujours appliqué. Il s'agit donc d'une constante à travers l'ensemble des 24 scénarios considérés.

Le critère 4

Le critère 4 concerne la fiabilité des liens. L'idée ici est de fixer différents niveaux de fiabilité pour examiner leur impact sur nos questions de recherche. Plus le niveau sera bas et plus le nombre de liens acceptés sera élevé. Un tel critère aura manifestement un effet sur la détermination du nombre de personnes qui ont à la fois dans SIDIS et CJCS. Nous allons pour conduire notre explication sur la fiabilité utiliser l'expression de « fenêtre » : plus la fenêtre est large, plus on accepte des liens, et plus le niveau de fiabilité sera bas.

Qu'entend-on par fiabilité ? Il s'agit de la capacité à associer aux personnes les enregistrements qui leur appartiennent et non les enregistrements qui ne leur appartiennent pas. Le fait d'échouer à associer un enregistrement à une personne bien qu'il lui appartienne est ce qu'on appelle un faux négatif. Le fait d'associer à une personne un enregistrement qui ne lui appartient pas est ce qu'on

appelle un faux positif. Dans le cadre de cet exercice, on souhaite naturellement diminuer et le nombre de faux négatifs et le nombre de faux positifs.

Nous illustrons la problématique de la fiabilité en examinant trois fenêtres de paramètres relatifs à la fiabilité des liens (voir Tableau 15). Chaque fenêtre est associée à un seuil de poids des preuves (MAX(S)) et un seuil de similarité (SIM) qu'il s'agit d'atteindre pour accepter un lien. Plus ces seuils seront élevés, plus le nombre de liens qui satisfont ces seuils sera réduit.

Tableau 15 – Trois fenêtres de paramètres relatifs à la fiabilité des liens

	Type de fenêtre	Symbole	Seuil de poids MAX(S)	Seuil de similarité SIM
1	Large	L	20	70%
2	Moyenne	M	25	80%
3	Étroite	E	30	90%

La fenêtre large [L] correspond à un seuil de poids de minimum 20 et de similarité de minimum 70%. Cela signifie qu'il faut une quantité de preuves (MAX(S)) égale ou supérieure à 30 et une similarité (SIM) égale ou supérieure à 70% pour accepter un lien. On notera qu'ici, 70% correspond à un score de minimum 14/20 (mais d'autres scores seront acceptables donc[29]). La fenêtre large [L] est la moins exigeante des fenêtres dans le sens où elle devrait correspondre au nombre de connexions entre enregistrements le plus élevé des trois fenêtres, au nombre de faux positifs le plus élevé et au nombre de faux négatifs le moins élevé.

La fenêtre moyenne [M] correspond à un seuil de poids de minimum 25 et de similarité de minimum 80% (ce qui revient à un score de minimum 20/25). La fenêtre moyenne [M] est un peu plus exigeante que la fenêtre large [L] dans le sens où elle devrait déboucher sur un nombre de connexions entre enregistrements moins élevé, et un nombre de faux positifs moins élevé (si elle inclut des mauvais liens d'intégration). En revanche son nombre de faux négatifs pourrait être plus élevé (si elle exclut des bons liens d'intégration).

La fenêtre étroite [E] correspond à un seuil de poids de minimum 30 et de similarité de minimum 90% (ce qui revient à un score de minimum 27/30). La fenêtre étroite [E] est plus exigeante que la fenêtre [M] dans le sens où elle devrait déboucher sur un nombre de connexions entre enregistrements moins élevé, et un nombre de faux positifs moins élevé. En revanche son nombre de faux négatifs pourrait être plus élevé, si elle exclut des bons liens d'intégration.

On comprend que diminuer une mesure de fiabilité (e.g., nombre de faux positifs) peut entraîner l'augmentation de l'autre mesure de fiabilité (i.e., nombre de faux négatifs), et vice versa. Mais comment sait-on quelle est la bonne fenêtre ? Quelle est la fenêtre optimale qui va minimiser le nombre de faux positifs et le nombre de faux négatifs ?

Il n'est pas facile de répondre à ces questions. Tout d'abord, nous ne disposons que des données présentes dans la base de données pour effectuer cette évaluation. Autrement dit, nous ne disposons pas de données extérieures permettant de la corroborer. Or les données disponibles ne contiennent

[29] Par exemple, $\frac{21}{30} = \frac{28}{40} = 70\%$ ou $\frac{15}{20} = 75\%$ ou encore $\frac{16}{20} = \frac{24}{30} = \frac{32}{40} = 80\%$.

pas de « lien vrai » (sinon nous n'aurions pas à effectuer cet exercice d'intégration en premier lieu). Si un tel lien était disponible, nous pourrions déterminer via un algorithme procédant par essai et erreur, quels seuils de poids et de similarité permettent le mieux de créer de bons liens et de ne pas créer de mauvais (i.e., augmenter le nombre de vrais positifs et diminuer le nombre de faux négatifs).

Ensuite, même s'il était possible de solliciter le concours d'une aide extérieure pour vérifier chaque cas (e.g., aller contrôler l'identité d'une personne en obtenant manuellement de l'information en provenance d'une source de données tierce), la grande quantité de données rendrait difficile d'effectuer une telle vérification.

Par conséquent, dans la présente recherche, nous n'avons pas cherché à estimer le nombre de faux positifs et le nombre de faux négatifs. Nous nous sommes contentés de développer une approche pragmatique visant à proposer différentes configurations de paramètres raisonnables, basées sur notre compréhension du domaine. Nous avons ensuite effectué des sondages manuels des liens, afin de détecter d'éventuels cas problématiques. L'exercice a consisté à observer un nombre limité de liens établis avec une variété de niveaux de poids et de similarité, en examinant les données personnelles (nom, prénom, date de naissance, etc.) des enregistrements ainsi reliés. Le but était de détecter les cas douteux (i.e., d'enregistrements supposément liés mais appartenant à des personnes différentes) et noter les seuils de poids et de similarité à partir desquels on cessait d'observer de tels cas.

En effectuant de tels sondages, nous avons estimé qu'un seuil de poids minimal de 25 et de similarité minimale de 80% (i.e., les seuils de la fenêtre [M]) consistait une bonne hypothèse de travail pour considérer un lien comme fiable. Il s'agit véritablement d'une hypothèse de travail : il n'est en effet pas certain que nous ne pourrions pas découvrir de cas douteux à l'avenir qui correspondent pourtant à ces seuils (i.e., des faux positifs). Par ailleurs, il est fort probable que nous ayons exclu des enregistrements qui mériteraient d'être liés, bien que leurs valeurs soient en-deçà de ces seuils (c'est-à-dire qu'il est fort probable que nous ayons produit des faux négatifs).

Quoi qu'il en soit, un tel choix de paramètres (fenêtre [M]) ne manquera pas d'influencer le nombre de personnes qui seront identifiées ainsi que certaines des statistiques qui seront établies à leur sujet (e.g., le nombre de personnes de SIDIS qui sont également dans CJCS). Afin de garder à l'esprit que ce choix est un choix méthodologique particulier qui exerce une influence sur les résultats, nous avons décidé de prendre en compte également les fenêtres [L] et [E], ainsi que d'autres critères (voir Tableau 16), expliqués ci-dessous.

Tableau 16 – Six fenêtres de paramètres relatifs à la fiabilité des liens

	Type de fenêtre	Symbole	Seuil de poids MAX(S)	Seuil de similarité SIM	Autre considération
1	Large	L	20	70%	/
2	Moyenne	M	25	80%	/
3	Étroite	E	30	90%	/
4	Moyenne exigeante	ME	27	80%	/
5	RRN	RRN	/	/	Avoir le RRN en commun
6	Moyenne exigeante ou RRN	ME/RRN	27	80%	Avoir le RRN en commun

Pour observer l'impact d'une modification minime d'un paramètre, on a défini la fenêtre moyenne exigeante [ME] qui est plus exigeante que la fenêtre moyenne [M] dans la mesure où le seuil de la preuve est augmenté de 25 à 27. Le seuil de similarité SIM, quant à lui, ne change pas.

Imaginons que l'on hésite à utiliser les fenêtres proposées jusqu'ici ([L], [M], [E], [ME]) et que l'on préfère se fier à des liens dans lesquels on estime avoir une plus grande confiance, disons des liens établis uniquement sur la base du RRN (et rien d'autre). Une telle stratégie correspond à la fenêtre [RRN] qui établit un lien dès qu'il y a un RRN en commun et peu importe le degré de similarité entre les personnes (celui pouvant dès lors varier librement entre 0% et 100%). Un tel scénario n'est pas parfait et pourrait induire des faux positifs dans la mesure où un RRN pourrait être attribué à tort à un enregistrement[30].

Par ailleurs, la fenêtre [RRN] entraîne vraisemblablement un grand nombre de faux négatifs, c'est-à-dire des cas où deux enregistrements pourraient être assignés à la même personne, bien qu'un RRN ne soit pas disponible pour chacun d'eux. Autrement dit, il y a parmi l'ensemble des personnes considérées des personnes qui correspondent à des enregistrements isolés alors qu'ils auraient pu être réunis. On pourrait dès lors souhaiter écarter ces enregistrements indûment isolés. Une manière de procéder pourrait être de se concentrer uniquement sur les personnes pourvues d'un RRN (que ce dernier ait ou non permis d'établir une connexion).

Enfin, avec la fenêtre [ME/RRN] on retient un lien s'il satisfait les conditions de la fenêtre [ME] ou les conditions de la fenêtre [RRN], ou les deux. C'est une manière de diminuer le nombre de faux négatifs évoqué précédemment.

En proposant ces six fenêtres, on souhaite se donner l'opportunité d'inspecter lesquelles correspondent aux nombres de faux positifs et de faux négatifs les plus bas possibles. Toutefois, comme nous l'avons déjà dit, nous n'avons pas développé de mesures du nombre de faux positifs et du nombre de faux négatifs.

Dans le futur, il sera important d'établir une stratégie de validation des liens découverts ou de sélection des liens, qui permette de se faire une meilleure idée de la nature des liens établis en fonction de la méthodologie utilisée. Nous nous bornons en effet ici à fixer des seuils, sans nous inquiéter de la nature des liens. On pourrait par exemple différencier les liens où les informations principales telles que le prénom, le nom et la date de naissance correspondent entre les deux enregistrements des autres types de liens (par exemple, les cas où le nom et la date de naissance correspondent ainsi que la nationalité et le genre mais pas le prénom).

Le critère 5

Le critère 5 concerne le statut des dossiers de CJCS. Au sein de CJCS, les dossiers ont un statut qui permet de savoir s'ils sont actifs ou inactifs. Les dossiers inactifs sont les dossiers dont le statut est « effacé » ou « fusionné ». Les dossiers « effacés » peuvent contenir des éléments inexacts ou qui ne

[30] Certaines de ces erreurs pourraient peut-être être évitées en vérifiant la qualité du RRN (e.g., en comparant le RRN et la date de naissance). Une telle vérification n'a pas été effectuée dans le présent exercice.

sont plus pertinents, tandis que les éléments des dossiers « fusionnés » ont été en principe repris dans le dossier actif correspondant (voir Huynen, Jeuniaux, et al., 2024).

L'information contenue dans les dossiers est utilisée dans la procédure d'intégration (à l'étape 3), qu'ils soient actifs ou inactifs. En effet, bien que les dossiers inactifs présentent manifestement moins d'intérêt que les dossiers actifs aux fins de l'intégration, ils pourraient contenir de l'information utile pour effectuer de la liaison (pensons par exemple au fait qu'un enregistrement soit corrigé dans une source A mais pas dans la source B ; disposer de l'information non corrigée de A permettra de lier l'enregistrement à B).

Quel impact ce choix a-t-il sur la création des nœuds de personnes ? Pour répondre à cette question, on considérera deux types de scénarios : ceux qui comprennent tous les dossiers (actifs ou inactifs) et ceux qui ne comprennent que les dossiers actifs. On pourra ainsi mesurer l'impact du fait d'inclure également des dossiers inactifs. Il y a à la base 3.860.989 nœuds de dossiers, et parmi eux 88.212 qui ne sont pas actifs (2% du total). Exclure ces dossiers reviendrait donc à conserver 98% des enregistrements de dossiers. Les exclure a-t-il un impact ? C'est ce que nous verrons.

Le critère 6

Le critère 6 vise à tester l'impact des liens d'intégration créés à l'intérieur des données sources. On a en effet des liens créés à l'intérieur des données sources (i.e., intra-SIDIS et intra-CJCS) et des liens créés entre les données sources (i.e., inter-SIDIS-CJCS).

Ces liens « intra » sont utiles afin de détecter d'éventuels enregistrements dupliqués. Sans les liens « intra », on pourrait toutefois détecter des enregistrements dupliqués de manière indirecte via les liens « inter » entre SIDIS et CJCS (e.g., un nœud de CJCS connecté à deux nœuds de SIDIS, ou un nœud de SIDIS connecté à deux nœuds de CJCS). Ces liens « intra » sont-ils superflus pour autant ? Pour tester cela, nous avons imaginé des scénarios avec et sans les liens « intra ».

24 scénarios

Les trois premiers critères sont des constantes méthodologiques. En revanche, les trois derniers critères envisagent des alternatives. Le critère 4 envisage six alternatives, le critère 5 envisage deux alternatives et le critère 6 envisage lui aussi deux alternatives. Par conséquent, nous envisageons en tout 24 scénarios méthodologiques différents (voir Tableau 17)[31].

Tableau 17 – Les vingt-quatre scénarios méthodologiques utilisés pour créer des nœuds de personne

	Type de dossiers CJCS	Liens inter ou intra	L	M	ME/RRN	ME	E	RRN
1	actif & inactif	inter & intra	1	2	3	4	5	6
2	actif & inactif	inter	7	8	9	10	11	12
3	actif	inter & intra	13	14	15	16	17	18
4	actif	inter	19	20	21	22	23	24

[31] (1 x 1 x 1 x 6 x 2 x 2) = 24

Pour chacun de ces 24 scénarios, nous avons une manière différente de créer des sous-ensembles et donc des nœuds de personnes différents. Chaque scénario de connectivité correspond à un scénario méthodologique particulier. Ces scénarios sont en quelque sorte des mondes virtuels séparés dans lesquels évoluent des personnes distinctes.

Enfin, techniquement parlant, ces 24 ensembles distincts de personnes se manifestent au sein de l'IHD par la création de 24 types de nœuds « personne » : des personnes appartenant au scénario 1, des personnes appartenant au scénario 2, etc., jusqu'au aux personnes appartenant au scénario 24 (voir Figure 12).

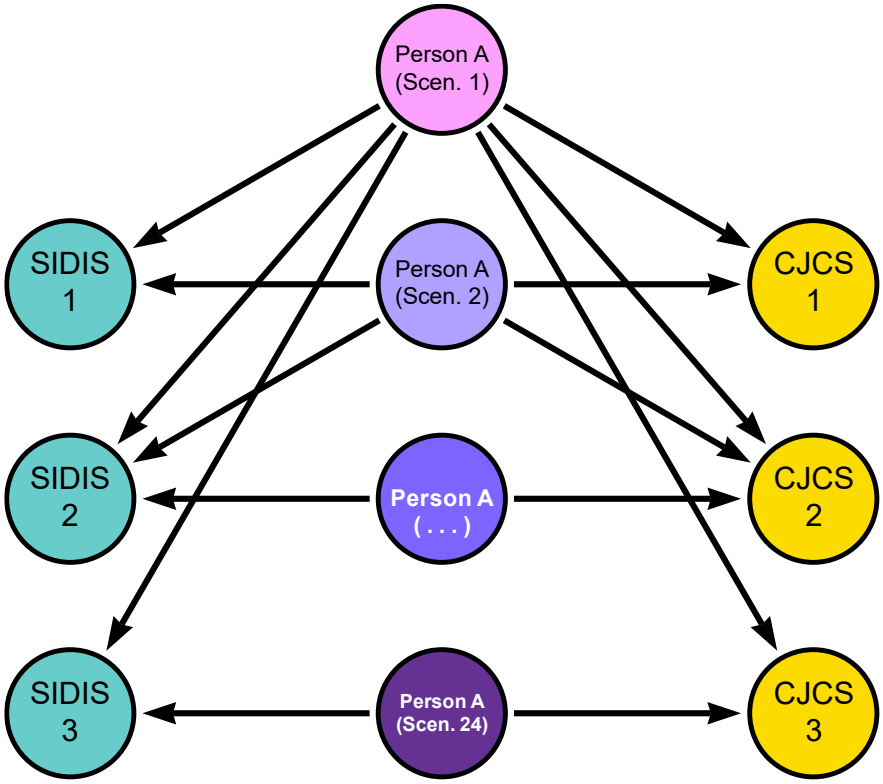


Figure 12 – Exemple fictif : création de nœuds de personnes selon différents scénarios

Comme on peut le voir dans la Figure 12, la personne A du scénario 1 est reliée à plus d'enregistrements que la personne A du scénario 24. Ces deux personnes A n'ont rien à voir l'une avec l'autre car elles appartiennent à des scénarios distincts (i.e., des mondes séparés).

2.6.5. Détails complémentaires

Une version préliminaire de la procédure de création de l'IHD a été mise au point en lien étroit avec la base de données qui a été développée dans le projet FAR (Jeuniaux et al. 2022). Ce travail étant fortement lié aux besoins du projet FAR, il ne prenait toutefois pas en compte l'ensemble des besoins du projet IIHA. Pour satisfaire l'ensemble de ces besoins, une nouvelle version de la procédure de création de l'IHD a été réalisée.

Par rapport au système développé dans FAR, pour le projet IIHA, l'IHD a fait l'objet de cinq types d'innovation.

Premièrement, son esprit est différent. Elle n'est pas strictement dévolue à un certain type d'analyse (e.g., analyse sur les carrières criminelles), mais entend assurer l'archivage des collections digitales de l'État dans le but de leur analyse ultérieure. Elle préserve donc les données et permet leur exploitation pour tout type d'usage (e.g., qui peut aussi bien concerner l'analyse des carrières criminelles que des analyses sur le fonctionnement des prisons).

Deuxièmement, le logiciel écrit pour le projet FAR a été totalement réécrit de manière à faciliter son usage futur dans d'autre projets. Il offre notamment un système permettant de décrire les données en vue de leur transformation en graphe, leur importation dans Neo4j et leur exploitation au sein du graphe ainsi formé. Un tel système permet d'exploiter plus aisément de nouvelles données. Par exemple, une nouvelle extraction de données de CJCS pourrait être reçue. Si la structure de CJCS devait avoir changé entretemps, il suffirait d'adapter la description des données pour les transformer en graphe et les importer dans Neo4j. De façon similaire, d'autres sources de données pourraient être ajoutées (e.g., de SIDIS-suite, des Maisons de justice).

Troisièmement, toutes les données disponibles qui ont été extraites de CJCS et de SIDIS ont été stockées dans le système nouvellement créé (sans qu'aucune sélection n'ait été effectuée en amont du processus de création de l'IHD, comme ça avait été le cas dans le cadre du projet FAR). Cette étape vise à satisfaire l'objectif de préserver et faciliter l'exploitation des Collections de l'État.

Quatrièmement, les modifications inhérentes à la création du graphe mises à part, aucune transformation préalable des données n'a été apportée avant de les stocker dans le nouveau système. Par ailleurs, une attention particulière a été apportée à distinguer les données brutes des informations supplémentaires résultant de l'analyse de données, ce, à nouveau, afin de respecter l'objectif de préservation des données.

Cinquièmement, la procédure d'intégration des données a été améliorée par l'adjonction de nouvelles méthodes d'intégration et la réalisation d'analyses plus éclairantes par rapport à leur valeur ajoutée respective. Par ailleurs, la procédure d'intégration a été repensée via la création d'un nouveau type nœud représentant la personne (une personne pouvant être lié à un ou plusieurs enregistrements) de façon à faciliter l'exploitation des données relatives aux personnes ou à faciliter les analyses reposant sur une intégration des données au niveau de la personne.

Pour terminer, les spécifications techniques de l'environnement de travail utilisé pour développer l'IHD sont données dans le Tableau 18.

Tableau 18 – Environnement de travail utilisé pour développer l'IHD

PROPRIÉTÉ	VALEUR
ORDINATEUR	TUXEDO Book XUX7 Gen 13. Intel Core i9-11900 K, 128 GB de RAM (4 x 32 GB) 3200 Mhz CL22 Samsung, NVIDIA GeForce RTX 3080 15 GB
DISQUE DUR EXTERNE	SSD de type iStorage diskAshur PRO2 (1 TB), norme AES à 256 bits, mode de chiffrement par bloc XTS.
LINUX	TUXEDO / UBUNTU 22.04 LTS
GNU BASH	5.1.16(1)-release
ANACONDA	2023.07
PYTHON	3.11.3
JUPYTER	6.5.4
NEO4J	4.4.4
AWESOME PROCEDURES ON	4.4.0.3
GRAPH DATA SCIENCE (GDS)	1.8.3
AZUL PLATFORM CORE (ZULU)	11.54.23
JAVA RUNTIME ENVIRONMENT	11.0.14
VISUAL STUDIO CODE	1.80.1

3. Résultats

3.1. Étape 1 : les nœuds et relations de base

L'ensemble des données disponibles du Casier Judiciaire Central (CJCS) et de SIDIS-greffe (SIDIS) ont été traitées et converties dans un graphe au sein de Neo4j.

Dans CJCS, les enregistrements de personnes sont stockés dans la table DOSSIER, tandis que dans SIDIS, les enregistrements de personnes sont stockés dans la table SIGNALETIEKEN. Dans la représentation en graphe, chaque enregistrement correspond à un nœud.

Nous avons 3.860.989 nœuds d'enregistrement de personnes dans CJCS et 365.401 nœuds de fiches signalétiques dans SIDIS, et par conséquent un total de 4.226.390 nœuds d'enregistrements de personnes.

Il s'agira d'appliquer la procédure d'intégration à ces 4.226.390 nœuds afin de créer les nœuds de personnes correspondants. Parmi ces nœuds de personnes combien seront associés aussi bien à SIDIS qu'à CJCS ? Puisque nous avons appliqué 24 scénarios méthodologiques à la fin de la procédure d'intégration (à l'étape 4), nous aurons 24 réponses à cette question. Ainsi en sera-t-il de toute autre question que nous nous poserons (e.g., Combien de personnes se retrouvent uniquement dans SIDIS ?).

Au sein du graphe, chaque nœud est relié à un ou plusieurs autres nœuds par des relations particulières qui existent en vertu de la structure des données de départ. Dans les deux sections qui suivent, nous nous penchons sur la structure des données de CJCS et SIDIS.

3.1.1. CJCS

La modélisation des données de CJCS en graphe, c'est-à-dire la logique sous-jacente à la représentation en termes de nœuds et de relations donne lieu à la Figure 13 ci-dessous. Il s'agit d'une représentation simplifiée de la structure des données où, par souci de clarté, certains nœuds n'ont pas été représentés[32].

Par exemple, dans la Figure 13, un nœud de bulletin de condamnation (BULLETIN) indiquera quel est son nœud de dossier (DOSSIER), grâce à la relation qui a pour origine le nœud de bulletin et pour destination le nœud de dossier[33]. Un nœud de dossier pourra ainsi être associé à un ou plusieurs nœuds de bulletin de condamnation, chaque nœud de bulletin pointant vers lui.

En aucun cas un nœud de dossier ne sera lié à une décision de condamnation, car cela contreviendrait à la structure des données de départ. En effet, dans CJCS, une décision de condamnation est liée à un

[32] Par exemple, des nœuds sur le statut des dossiers, ou le statut des bulletins. Cette information est toutefois bien présente dans l'IHD.

[33] La relation est représentée graphiquement par une flèche qui part du nœud de bulletin et pointe vers le nœud de dossier.

bulletin de condamnation et jamais à un dossier directement. Par conséquent, dans la Figure 13, il n'y a pas de relation entre le nœud de décision et le nœud de dossier.

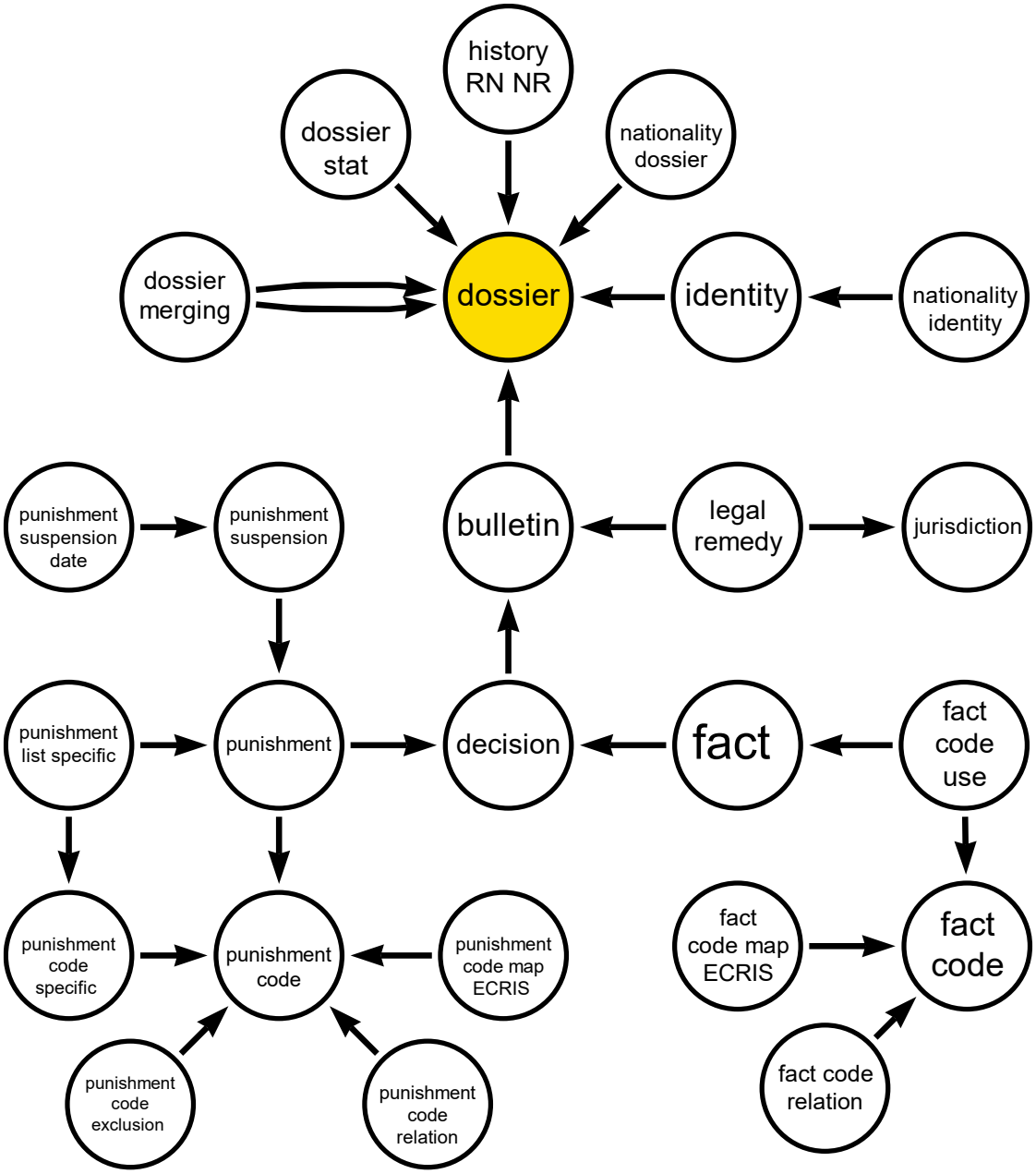


Figure 13 – Schéma graphique de CJCS – version simplifiée

3.1.2. SIDIS

La modélisation des données de SIDIS en graphe donne lieu à la Figure 14 ci-dessous. Là encore, il s'agit d'une représentation simplifiée de la structure des données où, par souci de clarté, certains nœuds ne sont pas représentés.

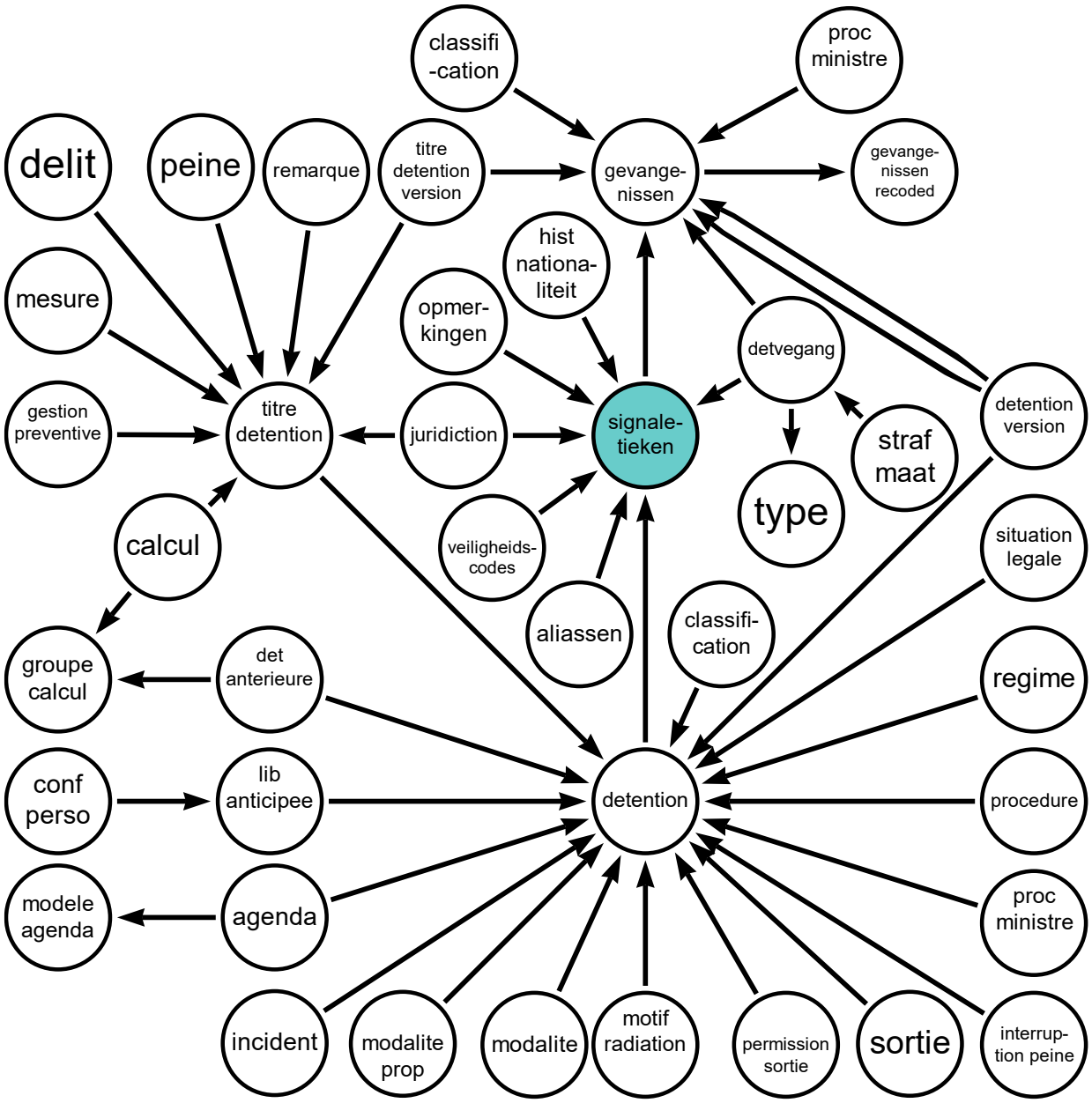


Figure 14 – Schéma graphique de SIDIS – version simplifiée

Par exemple, dans la Figure 14, un nœud de détention indiquera quel est son nœud de fiche signalétique (SIGNALETIEKEN), grâce à la relation qui a pour origine le nœud de détention (DETENTION) et pour destination le nœud de fiche signalétique. Un nœud de fiche signalétique pourra ainsi être associé à un ou plusieurs nœuds de bulletin de détention, chaque nœud de détention pointant vers lui.

Par ailleurs, on trouve dans la modélisation exprimée par la Figure 14, deux types de nœuds dont la dénomination n'apparaît pas explicitement dans les tables de SIDIS : la version de la détention (DETENTION VERSION), et la version du titre de détention (TITRE DETENTION VERSION). Cette innovation par rapport aux données de départ a été rendu nécessaire par la nature du graphe en Neo4j. Pourquoi cela ? Dans SIDIS, on a uniquement deux tables à ce sujet : DETENTION et TITRE DETENTION. Toutefois, la table DETENTION se réfère à des versions de détention (i.e., plusieurs versions d'une même détention). Pareillement, la table TITRE DETENTION se réfère à des versions du titre de détention (i.e., plusieurs versions d'un même titre de détention). Or un nœud dans le graphe

doit pouvoir être identifié de manière unique. La manière de procéder a donc été de rendre explicite l'existence des versions de détention (au sein de nœuds appelés DETENTION VERSION) et des versions de titres de détention (au sein de nœuds appelés TITRE DETENTION VERSION), et d'ajouter les détentions et titres de détention proprement dits dans des nœuds DETENTION ET TITRE DETENTION.

Enfin, comme on le voit dans la Figure 14, le mouvement du détenu au cours de sa détention (DETGEVANG) est lié à la fiche signalétique (SIGNALETIEKEN) et non pas à la détention (DETENTION) ou la version de la détention (DETENTION VERSION).

3.2. Étape 2 : les nœuds d'intégration

Vingt types de nœuds d'intégration ont été créés pour pouvoir relier et comparer les enregistrements entre eux selon la procédure mentionnée dans la section sur la méthodologie (voir Tableau 19). Dans le tableau, le nombre de nœuds de chaque type est donné, ainsi que le nombre de nœuds de CJCS et le nombre de nœuds de SIDIS reliés à ces nœuds d'intégration.

Par exemple, à la première ligne du Tableau 19, on peut lire qu'il y a 3.433.080 nœuds d'intégration de type « RRN » différents. Autrement dit, il y a 3.433.080 numéros de Registre National différents. Ensuite, parmi les nœuds d'enregistrements de personnes dans CJCS, il y en a 3.437.593 qui ont un RRN (i.e., 89,03 % des 3.860.989 dossiers de CJCS).

Et parmi les nœuds d'enregistrements de personnes dans SIDIS, il y en a 56.220 qui en ont un RRN (i.e., 15,39 % des 365.401 fiches signalétiques de SIDIS). Par conséquent, tous les enregistrements de personnes n'ont pas un RRN, mais la couverture est bien meilleure dans CJCS (89,03 %) que dans SIDIS (15,39 %). Le fait qu'il y ait plus de nœuds de CJCS liés à un RRN (3.437.593) que de RRNs différents (3.433.080) indique qu'un certain nombre de nœuds de CJCS pointent vers le même nœud de RRN.

À la ligne 2 du Tableau 19, on voit qu'on a trois genres différents. Il s'agit de : homme, femme, indéfini. À la ligne 3, on voit qu'on a 237 nationalités.

À la ligne 4, on apprend qu'on a 547.238 nœuds d'intégration « prénom » différents, et que 3.857.891 enregistrements de personnes dans CJCS sont associés à l'un de ces nœuds.

À la ligne 5, on voit que le nombre de combinaisons de trois premières lettres du prénom est assez réduit : 9.914. Par ailleurs, seuls 3.809.870 enregistrements de personnes dans CJCS sont associés à une combinaison des trois premières lettres du prénom, et non pas 3.857.891. La raison en est que dans un certain nombre de cas, le prénom n'était constitué que par une ou deux lettres, celles-ci ne pouvant pas constituer un ensemble de trois lettres. On observe la même chose dans le cas de SIDIS.

À la ligne 6, on voit que le nombre de représentations phonétiques du prénom est plus élevé que le nombre de combinaisons de trois premières lettres du prénom : 43.372. Cette fois-ci, il y a autant d'enregistrements de personnes de CJCS qui sont associés à une représentation phonétique du prénom que d'enregistrements de personnes CJCS associés à un prénom : 3.857.891. On observe la même chose dans le cas de SIDIS.

Tableau 19 – Fréquences relatives aux nœuds d'intégration

	Nœuds d'intégration	Fréquence	Nœuds de CJCS concernés		Nœuds de SIDIS concernés	
1	RRN	3.433.080	3.437.593	89,03 %	56.220	15,39 %
2	Genre	3	3.860.989	100,00 %	365.401	100,00 %
3	Nationalité	237	3.757.938	97,33 %	365.224	99,95 %
4	Prénom	547.238	3.857.891	99,92 %	363.145	99,38 %
5	Trois premières lettres du prénom	9.914	3.809.870	98,68 %	362.991	99,34 %
6	Représentation phonétique du prénom	43.372	3.857.891	99,92 %	363.145	99,38 %
7	Nom de famille	547.238	3.860.984	100,00 %	365.293	99,97 %
8	Trois premières lettres du nom de famille	9.914	3.858.148	99,93 %	364.773	99,83 %
9	Repr. phonétique du nom de famille	43.372	3.860.984	100,00 %	365.293	99,97 %
10	Date de naissance	36.613	3.785.900	98,06 %	364.757	99,82 %
11	Année de naissance	155	3.817.238	98,87 %	365.401	100,00 %
12	Mois de naissance	12	3.785.915	98,06 %	364.757	99,82 %
13	Jour de naissance	31	3.795.372	98,30 %	364.757	99,82 %
14	Pays de naissance	266	3.680.407	95,32 %	364.756	99,82 %
15	Lieu de naissance	53.461	2.349.432	60,85 %	363.460	99,47 %
16	Pays de résidence	218	3.419.386	88,56 %	156.580	42,85 %
17	Lieu de résidence	8.368	3.234.991	83,79 %	151.497	41,46 %
18	Code postal du lieu de résidence	1.854	3.178.901	82,33 %	125.998	34,48 %
19	Adresse du lieu de résidence	2.424.847	3.347.390	86,70 %	155.850	42,65 %
20	Date de jugement	25.791	3.073.343	79,60 %	365.094	99,92 %

Les lignes 7-9 concernent le nom de famille. Les mêmes phénomènes que ceux observés pour le prénom s'observent ici aussi. Tous les enregistrements de personnes qui sont associés à un nom de famille, sont associés à une représentation phonétique du nom de famille mais ils ne sont pas tous associés à une combinaison des trois premières lettres du nom de famille, car dans certains cas celui-ci n'était constitué que d'une ou deux lettres.

Il faut aussi noter qu'on a en apparence le même nombre de prénoms que de noms de famille (i.e., 547.240). Or ce n'est pas vrai. En réalité, on a 156.174 prénoms distincts et 435.061 noms de famille distincts. La raison qui fait qu'on obtient 547.240 nœuds d'intégration « prénom » et 547.240 nœuds d'intégration « nom » est que les prénoms et les noms de famille sont modélisés doublement à cause du procédé d'inversion décrit dans la section sur la méthodologie. Un tel dispositif fait qu'on a 547.240 nœuds qui peuvent à la fois jouer le rôle de prénom ou de nom.

Les lignes 10-13 concernent les date, année, mois et jour de naissance. Il n'y pas de correspondance directe entre le nombre d'enregistrements associés à ces différentes informations, car chaque information (date, année, mois ou jour) peut ou non être valide (au sens défini dans le Tableau 9). Par exemple, le numéro du mois pourrait n'être pas valide (i.e., ne pas être un nombre situé entre 1 et 12), mais bien le jour, et vice versa. Ainsi, dans CJCS, on a 3.817.238 enregistrements avec une année de naissance valide, 3.795.372 enregistrements avec un jour de naissance valide, 3.785.915 enregistrements avec un mois de naissance valide, et 3.785.900 enregistrements avec une date de naissance valide.

Pour terminer, à la ligne 20, on voit que la quasi-totalité des personnes de SIDIS ont une date de condamnation (99,92 %), mais que seules 8 personnes sur 10 dans CJCS (79,60 %) présentent une telle date.

Le fait qu'on n'ait pas 100% de date de condamnation encodées dans CJCS est à première vue étonnant puisque le but de CJCS est évidemment d'encoder les condamnations. La raison en est que 787.646 dossiers (20,4 %) stockés dans l'IHD ne sont pas associés à des bulletins de condamnation, et par conséquent, une date de jugement n'est pas disponible pour ces dossiers[34].

Le très haut taux de remplissage d'une date de jugement pour SIDIS est peut-être plus étonnant encore, car une telle information n'a en principe été encodée qu'après les années 2000. On devrait donc s'attendre à observer un taux inférieur. Le champ utilisé pour trouver cette date est le champ 'date_jugement' de la table JURIDICTION de SIDIS (voir Tableau 8), qui est lié à la fiche signalétique (voir Figure 14). Un travail subséquent d'analyse devra être conduit afin de déterminer s'il n'y a pas un meilleur moyen d'exploiter cette information.

Quoi qu'il en soit, la date de jugement n'est qu'une information parmi d'autres pour évaluer la similarité entre les enregistrements, or, comme on l'a vu dans la section 2.6.3, l'information sur la date de jugement n'a pas reçu un rôle prépondérant dans cette évaluation (elle ne joue aucun rôle dans la phase 1 de dégrossissage et dans la phase 2, elle n'a qu'un poids de 1).

[34] Dans le présent exercice, nous n'avons pas analysé les caractéristiques de ces dossiers. Il serait utile de le faire pour déterminer pourquoi il n'y a pas de bulletin correspondant, et éventuellement traiter ces dossiers de manière particulière.

3.3. Étape 3 : les liens d'intégration

La création des liens d'intégration a été effectuée pour les six méthodes envisagées (M^{RRN}, M^{JUG}, M^{TRI}, M^{PHO}, M^{TRI-i}, M^{PHO-i}), selon trois types d'orientation des liens : au sein de SIDIS seul (intra-SIDIS), au sein de CJCS seul (intra-CJCS) et entre SIDIS et CJCS (inter-SIDIS-CJCS).

3.3.1. Quantités de liens établis

Pour rappel, la création des liens se fait en général[35] en deux phases : d'abord une phase de dégrossissage puis une phase d'affinage. La phase 1 de dégrossissage consiste en une comparaison vectorielle rapide entre les enregistrements au moyen du coefficient de Jaccard. La phase 2 d'affinage consiste en une comparaison textuelle plus lente entre les enregistrements au moyen de la mesure de similarité textuelle de Jaro-Winkler. Elle est plus lente mais elle opère sur un nombre d'enregistrements plus réduit grâce au premier passage de la méthode de dégrossissage.

Phase 1 : dégrossissage

Plus de 5,7 millions de liens sont produits au dégrossissage. Comme on peut le voir dans la Figure 15, à cette étape, plus de liens sont créés au sein de CJCS (intra-CJCS) qu'au sein de SIDIS (intra-SIDIS) et le nombre de liens créés entre SIDIS et CJCS (inter-SIDIS-CJCS) est entre les deux. Environ 2,9 millions de liens sont créés intra-CJCS, 1,1 millions de liens intra-SIDIS et 1,7 millions de liens entre SIDIS et CJCS[36].

[35] C'est vrai pour toutes les méthodes, sauf pour la méthode M^{RRN}, pour laquelle on applique directement la phase 2 sans passer par la phase 1.

[36] Les chiffres précis sont disponibles dans le Tableau 24 reporté en annexe.

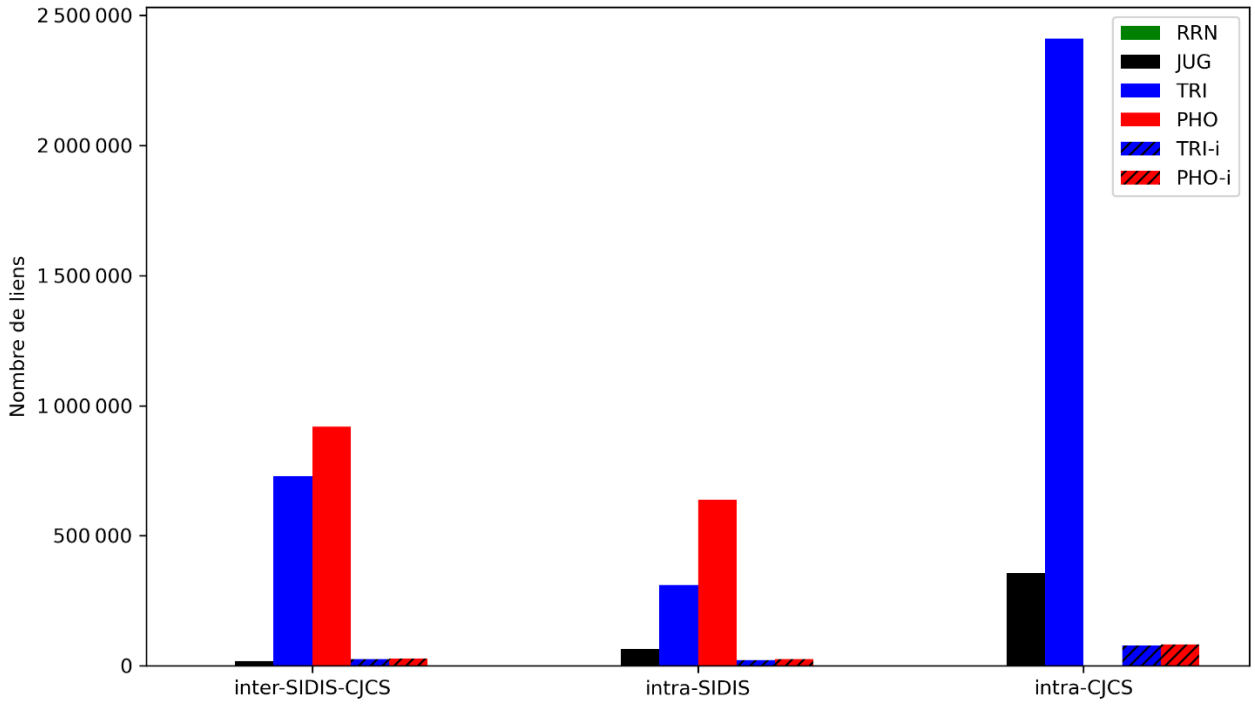


Figure 15 – Nombre de liens créés à l'étape de dégrossissage (selon l'orientation et la méthode)

La quantité de liens la plus importante est générée par la méthode M^{TRI} au sein de CJCS (environ 2,4 millions de liens). La méthode M^{PHO} n'a pas été appliquée jusqu'à son terme au sein de CJCS en raison du trop grand nombre de liens qu'elle s'apprêtait à créer et du temps d'exécution exorbitant que cela impliquait[37]. L'exécution de la méthode a donc été interrompue et le nombre de liens créés n'est pas disponible (d'où l'absence de barre dans le graphique). Le nombre de liens produits avec M^{PHO} (barre en rouge) apparaît en fait toujours supérieur au nombre de liens produits avec M^{TRI} (barre en bleu).

Ensuite, la méthode M^{JUG} produit environ 437.000 liens, tandis que la méthode M^{TRI-i} et M^{PHO-i} produisent 127.000 et 134.000 liens respectivement. Enfin, comme la méthode M^{RRN} n'est pas appliquée à cette étape, aucun nombre ne correspond à cette méthode dans le graphique.

Phase 2 : affinage

En phase 2, l'ensemble des liens produits en phase 1 sont passés en revue et « affinés ». Par ailleurs, des liens sont aussi ajoutés via la méthode M^{RRN}. À l'issue de la phase 2, on obtient environ 1,6 million de liens, soit environ 30% du nombre de liens produits en phase 1.

Comme on peut le voir dans la Figure 16, davantage de liens sont créés entre SIDIS et CJCS qu'au sein de SIDIS ou CJCS seuls. Quand on cherche à établir des liens entre SIDIS et CJCS, les méthodes M^{TRI} et M^{PHO} produisent le plus de liens (~ 490.000 et ~ 548.000 liens respectivement)[38]. Les méthodes M^{TRI-i} et

[37] Des chiffres sur les temps d'exécution sont disponibles en annexe, dans la section A.2.

[38] Les chiffres précis sont disponibles dans le Tableau 25 de l'annexe.

M^{PHO-i} produisent un nombre de liens beaucoup plus modeste (~ 20.000). Le nombre de liens établis avec M^{JUG} est lui aussi modeste avec environ 17.000 liens créés. La méthode M^{RRN} génère approximativement 50.000 liens.

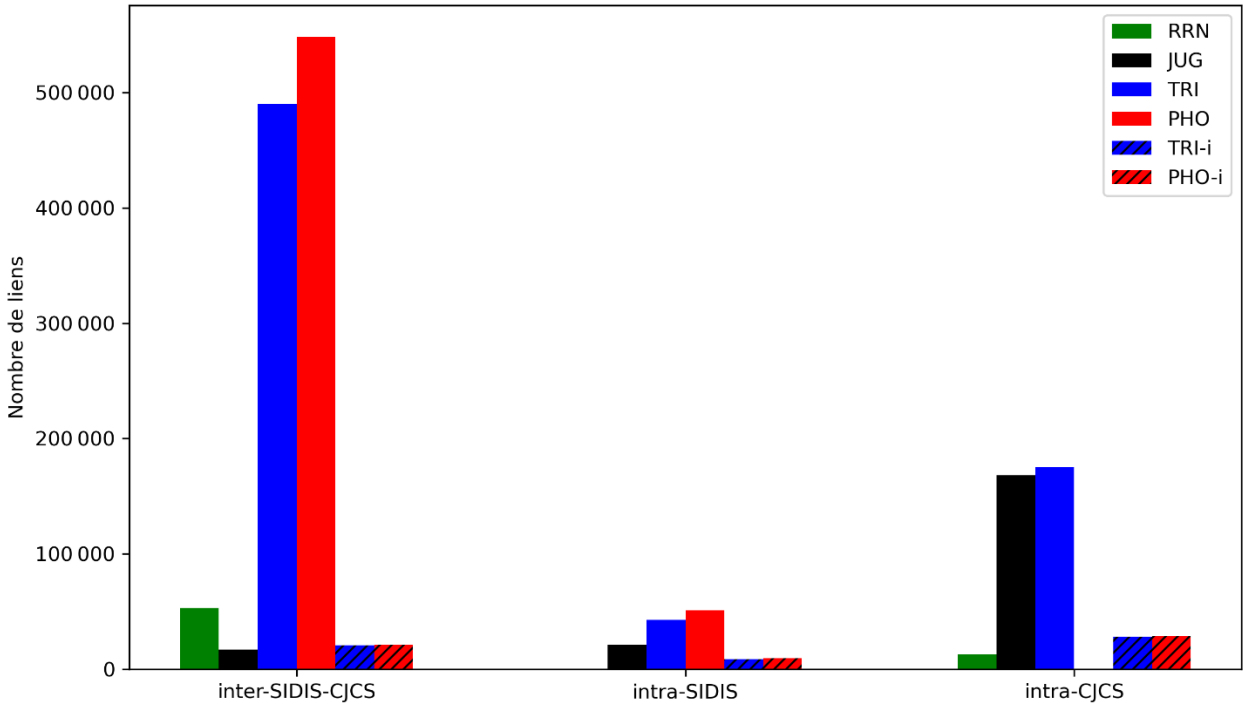


Figure 16 – Nombre de liens créés à l'étape d'affinage (selon l'orientation et la méthode)

Au total, on obtient environ 1,6 million de liens à l'issue de la phase 2. Cependant, tous ces liens ne sont pas équivalents entre eux. Certains sont en effet de nature douteuse, parce que leur niveau de similarité est trop faible, ou bien parce que le poids des preuves examinées est insuffisant, pour s'assurer qu'on a bien à faire aux mêmes personnes. Dans les sections qui suivent nous examinons le poids des preuves, ainsi que la similarité des liens, en nous focalisant dans notre explication sur l'orientation inter-SIDIS-CJCS, car c'est là que se situe le cœur de notre propos (i.e., la détection des individus qui possèdent des enregistrements aussi bien dans SIDIS que dans CJCS).

3.3.2. Le poids des preuves

Chaque lien établi entre deux enregistrements dispose d'un « poids » qui représente la quantité de preuves qui était disponible afin de juger de la similarité de ces enregistrements (i.e., MAX(S)).

Pour rendre compte de la distribution des poids, on a compté la fréquence de ces différents poids pour tous les liens établis que ce soit au sein de SIDIS seul, CJCS seul, ou entre SIDIS et CJCS[39]. Nous nous

[39] Ces données sont disponibles dans le Tableau 28, le Tableau 29 et le Tableau 30 de l'annexe à la section A.3.

focalisons ici sur les liens entre SIDIS et CJCS, et représentons les fréquences des poids dans un graphe de densité (voir Figure 17). Le poids est sur l'axe des X et la densité sur l'axe des Y.

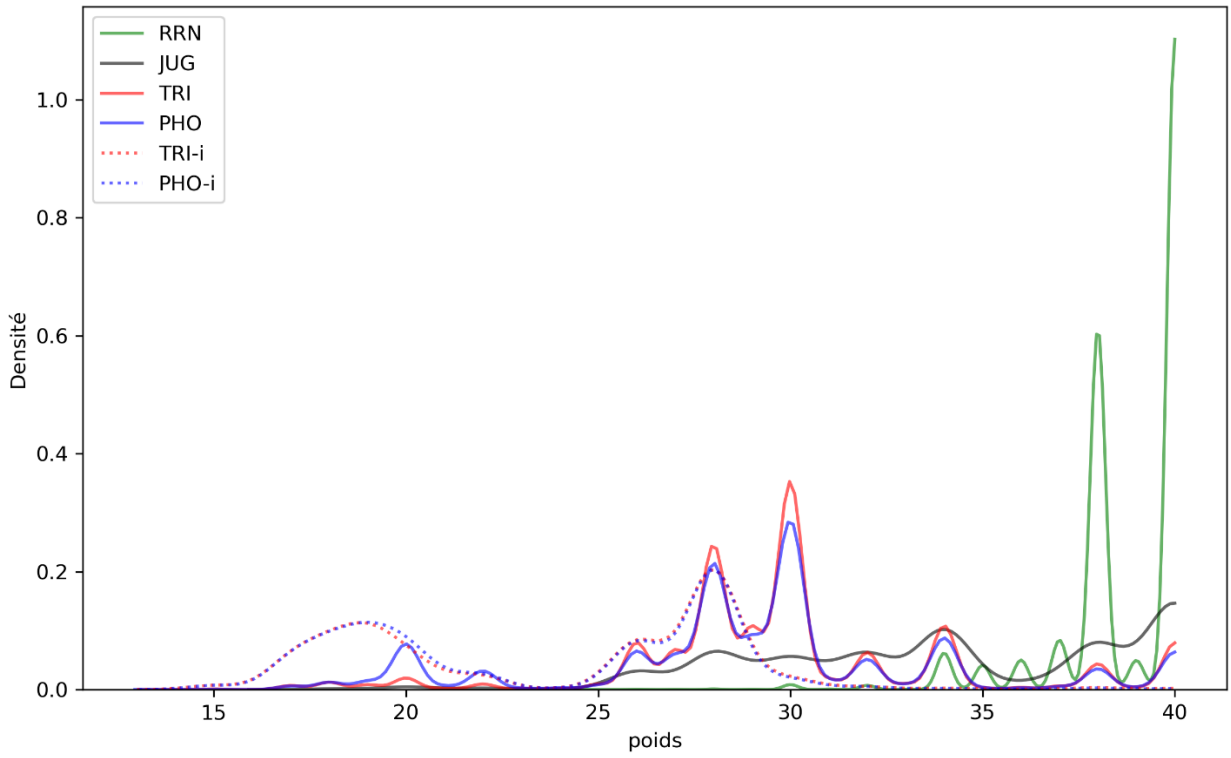


Figure 17 – Densité des poids (MAX(S)) des liens entre SIDIS et CJCS pour chacune des six méthodes

En inspectant la Figure 17, on observe que la méthode M^{RRN} concentre ses liens vers les poids les plus élevés (proche de 40), les méthodes M^{TRI} et M^{PHO} sont fort concentrées autour de 30 et en dessous, tandis que M^{TRI-i}, M^{PHO-i} sont fort concentrées entre 25 et 30, et entre 15 et 20. La méthode M^{JUG} concentre ses valeurs plus ou moins uniformément entre 25 et 40.

On peut conclure que M^{RRN} parvient à récupérer davantage de liens qui ont plus de poids que les autres méthodes. Son poids le plus petit est 25, mais le poids est en général beaucoup plus élevé. Autrement dit, si deux enregistrements ont le RRN en commun, ils ont généralement d'autres éléments de preuve susceptibles d'être comparés (tels que le prénom, nom, etc.). Par conséquent, une telle comparaison possède plus de poids.

On peut simplifier la réflexion visant la Figure 17 en calculant la proportion de liens, par exemple disposant d'un poids de 34 ou 37 (voir Tableau 20). Grâce à ces points de repères, on peut ordonner les méthodes selon leur capacité décroissante à rassembler du poids selon l'ordre suivant : 1) M^{RRN}, 2) M^{JUG}, 3) M^{TRI}, 4) M^{PHO}, 5) M^{TRI-i} et M^{PHO-i}.

Tableau 20 – Quantité de liens satisfaisant le critère de poids pour un lien entre SIDIS et CJCS selon les six méthodes

Critère de	M^{RRN}	M^{JUG}	M^{TRI}	M^{PHO}	M^{TRI-i}	M^{PHO-i}
Au moins 37	91%	41%	11%	9%	1%	1%
Au moins 34	99%	60%	20%	18%	1%	1%

Observe-t-on le même ordonnancement des méthodes quand on examine non pas le poids mais la similarité des liens ? Comme nous allons le voir ci-dessous, la réponse est : oui.

3.3.3. La similarité des liens

Chaque lien établi entre deux enregistrements dispose non seulement d'un poids, mais aussi du score de similarité (SIM) des enregistrements.

Pour rendre compte de la distribution de ces indices de similarité, on a compté la fréquence des valeurs en les plaçant dans 11 grandes catégories : 100%, au moins 90%, au moins 80%, etc. On a procédé à ce comptage pour tous les liens établis que ce soit au sein de SIDIS seul, CJCS seul ou entre SIDIS et CJCS[40]. Nous nous focalisons ici sur les liens établis entre SIDIS et CJCS, et représentons les fréquences des similarités via un graphe de densité (voir Figure 18). La similarité est sur l'axe des X et la densité sur l'axe des Y.

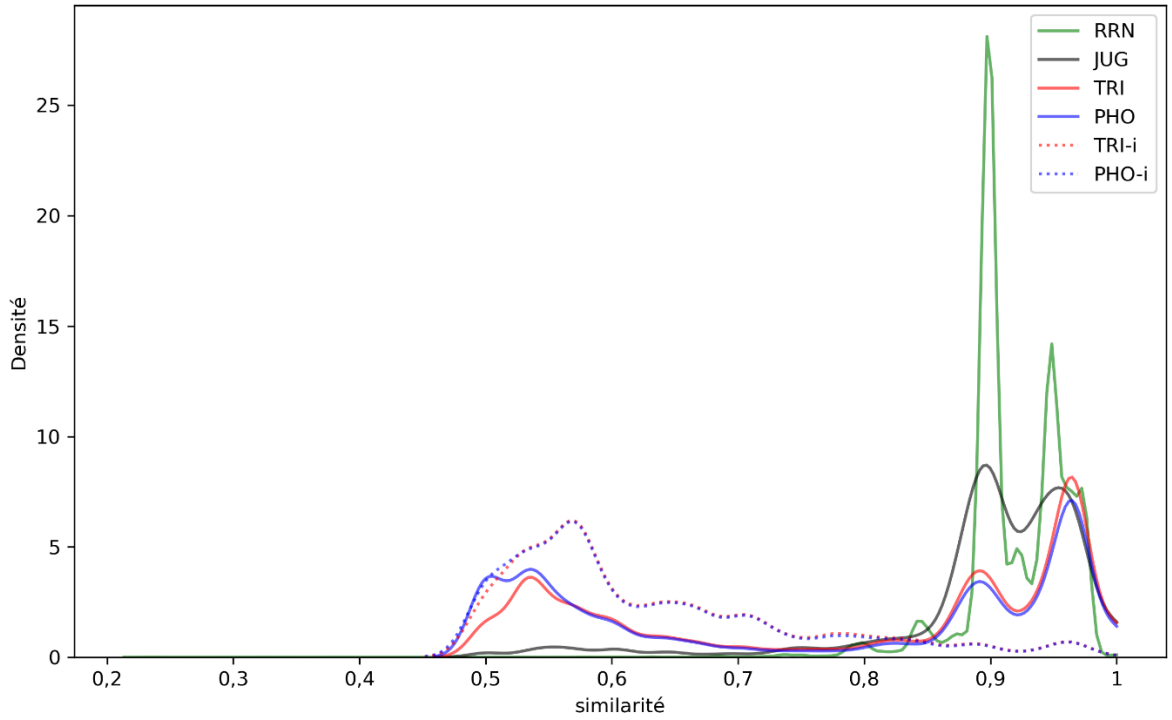


Figure 18 – Densité des similarités (SIM) des liens entre SIDIS et CJCS pour chacune des six méthodes

La majorité des liens établis grâce à M^{RRN} sont de similarité élevée (autour de 90%). Des liens de similarité supérieure à 90% sont plus rares. Le niveau de similarité est donc généralement imparfait. Autrement dit, certaines données tendent à être différentes[41], même lorsque les enregistrements partagent le même RRN.

[40] Ces données sont disponibles dans le Tableau 31, le Tableau 32 et le Tableau 34 de l'annexe à la section A.4.

[41] Un examen complémentaire portant sur ces différences devra être conduit à un moment ultérieur.

Quant à M^{TRI}, elle rassemble certes une grande quantité de liens de similarité très élevée (90% et plus), mais aussi une grande quantité de liens de similarité faible (entre 50% et 60%). Il en est de même pour M^{PHO}.

Comme on l'a fait précédemment, on peut simplifier la réflexion visant la Figure 17 en calculant la proportion de liens, par exemple avec au moins 80% de similarité ou 90% de similarité (voir Tableau 21).

Tableau 21 – Quantité de liens satisfaisant le critère de similarité – selon les six méthodes

Critère de	M^{RRN}	M^{JUG}	M^{TRI}	M^{PHO}	M^{TRI-i}	M^{PHO-i}
Au moins 90%	78%	65%	44%	39%	4%	4%
Au moins 80%	99%	91%	60%	53%	12%	12%

Par exemple, parmi les liens récupérés par la méthode M^{RRN}, 78% d'entre eux ont au moins une similarité de 90%, et 99% d'entre eux ont au moins une similarité de 80%. À l'autre bout du tableau, on voit que parmi les liens récupérés par la méthode M^{PHO-i}, 4% d'entre eux ont au moins une similarité de 90%, et 12% d'entre eux ont au moins une similarité de 80%.

Grâce à ces points de repères, on peut ordonner les méthodes selon leur capacité décroissante à rassembler de la similarité. L'ordre est le même que celui observé précédemment dans le Tableau 20.

Cela signifie-t-il que M^{PHO-i} est inutile comparé à M^{RRN} ? Non, on ne peut pas dire cela, car il se peut que M^{PHO-i} récupère des liens de bonne qualité que M^{RRN} ou d'autre méthode ne récupère pas. Comme nous le verrons dans la section 3.3.5 sur les « valeurs ajoutées », c'est effectivement le cas.

3.3.4. Seuils de poids et de similarité

Considérer le niveau du poids et de similarité isolément n'est toutefois pas suffisant : il faut pouvoir les appréhender simultanément. Pour qu'un lien soit considéré comme « suffisamment » fiable, il faut que « suffisamment » de preuves aient été examinées et qu'après examen de ces preuves, le niveau de similarité soit considéré comme « suffisant ». À partir de quel moment considère-t-on qu'une mesure soit suffisante ? À quel niveau de poids et de similarité peut-on considérer un lien comme acceptable, c'est-à-dire qui concerne effectivement deux enregistrements associés à la même personne ?

Comme nous l'avons expliqué dans la section méthodologique, il n'est pas facile de répondre à cette question, principalement parce que nous ne disposons pas de données externes pour corroborer ce jugement. Dans le présent exercice, nous nous sommes bornés à établir différentes valeurs seuil raisonnables (voir Tableau 16) et en fonction de ces différents seuils, examiner un échantillon des liens retenus afin d'y détecter d'éventuels cas douteux.

Les six fenêtres du Tableau 16 ont été appliquées pour définir six ensembles différents de liens. À titre illustratif, le nombre de liens conservés a été rapporté uniquement pour les fenêtres [L], [M] et [E] dans le Tableau 35, le Tableau 42 et le Tableau 48 de l'annexe.

Étant donné que les méthodes (M^{RRN}, M^{TRI}, M^{PHO}, etc.) diffèrent en termes de poids et de similarité de leurs liens, l'application des fenêtres n'aura pas le même impact sur ces différentes méthodes en termes du nombre de liens conservés. Par exemple, lorsqu'on applique la fenêtre [E] aux liens inter-

SIDIS-CJCS créés, 77% des 53.134 liens créés par la méthode M^{RRN} sont conservés, et seuls 28% des 490.086 liens créés par la méthode M^{TRI} sont conservés.

Par ailleurs le nombre de liens conservés variera selon que les liens ont été créés dans SIDIS seul, CJCS seul ou entre SIDIS et CJCS. Par exemple, avec la méthode M^{RRR}, lorsqu'on applique la fenêtre [E], si 77% des 53.134 liens inter-SIDIS-CJCS sont conservés, c'est 99% des 12.877 liens intra-CJCS qui sont conservés, et seulement 6% des 226 liens intra-SIDIS qui sont conservés.

3.3.5. Valeurs ajoutées des méthodes de liaison

Comme on le sait, deux enregistrements peuvent être liés ou pas par un lien d'intégration établi à l'étape 3. Quand ils sont liés, on dit qu'il y a « liaison ». Une liaison peut se manifester par un ou plusieurs liens, chaque lien pouvant être d'un type différent selon la méthode qui a été utilisée pour le tracer. Comme il y a six types de méthodes (e.g., M^{RRN}, M^{TRI}, M^{PHO}), une liaison peut être constituée par au plus six liens (voir Figure 7). Les méthodes sont-elles équivalentes ? Certaines sont-elles redondantes ? Ont-elles toutes une valeur ajoutée ? Par exemple, dans combien de situations a-t-on, pour une paire d'enregistrements, un lien basé sur M^{RRN} et non un lien basé sur M^{TRI} ? Dans combien de situations a-t-on un lien basé sur M^{TRI} et non sur tel ou tel autre type de lien ?

Si nous souhaitons pouvoir répondre à toutes les questions de ce genre, il nous faut pouvoir considérer tous les scénarios possibles où au moins un lien d'un certain type est établi. Puisqu'on a appliqué six méthodes pour établir des liens, le nombre de situations possibles est $(2^6 – 1) = (64 – 1) = 63$. On a donc 63 situations différentes (identifiées par leur numéro allant de 1 à 63). Chaque situation correspond donc à une liaison entre deux enregistrements qui se matérialise par un ou plusieurs liens (minimum 1 lien, maximum 6 liens).

Pour chacune des 63 situations de liaisons possibles, on compte le nombre de liaisons, c'est-à-dire le nombre de paires d'enregistrements liés l'un à l'autre. Nous comptons le nombre de liaisons, pour chacune des 63 situations, de manière différenciée selon qu'on cherche à lier les nœuds de SIDIS entre eux (intra-SIDIS), les nœuds de CJCS entre eux (intra-CJCS) ou les nœuds de SIDIS aux nœuds de CJCS (inter-SIDIS-CJCS).

Nous avons effectué cet exercice en filtrant les liens selon leur similarité et poids d'après les trois types de fenêtres ([L], [M], [E]). Ici cependant, par souci de brièveté et de clarté, nous rapportons essentiellement les résultats obtenus avec la fenêtre moyenne ([M]).

On a tout d'abord comptabilisé le nombre de liens que chaque méthode est capable de récupérer, quelle que soit la situation concernée. On voit que les méthodes M^{TRI} et M^{PHO} sont supérieures aux autres dans leur habileté à établir une grande quantité de liens (voir Tableau 22).

Tableau 22 – Quantité de liens que chaque méthode permet de récupérer (tous les scénarios et toutes les orientations confondues) – fenêtre moyenne [M]

	M^{TRI}	M^{PHO}	M^{RRN}	M^{JUG}	M^{TRI-i}	M^{PHO-i}
Au total	287.835	285.718	52.351	15.487	1.545	1.507
Isolément	3.066	1.063	254	65	44	32

Par ailleurs, si toute méthode est généralement accompagnée d'une ou plusieurs autres méthodes, elle apparaît parfois isolément. Cela signifie que si la méthode n'avait pas été employée, les liaisons correspondantes auraient été perdues (à tort ou à raison). Par exemple, $M^{PHO\text{-}i}$ récupère 32 liens à elle seule. Autrement dit, sans elle, ces 32 liens auraient été perdus.

Nous rapportons à présent les situations qui ont été observées. Sur les 63 situations possibles, seules 46 se sont manifestées. Pour faciliter leur exposition et se concentrer sur l'essentiel, on ne retient ici que les 17 situations les plus fréquentes (celles pour lesquelles on a pu comptabiliser au moins 100 liaisons). Le résultat de ce filtrage est le Tableau 23 ci-dessous[42].

Tableau 23 – Distribution des liens selon les six méthodes de liaison et les bases de données concernées (fenêtre de paramètres moyenne [M]) – les 17 premières situations rangées de la plus fréquente à la moins fréquente

	M^{JUG}	M^{RRN}	M^{TRI}	$M^{TRI\text{-}i}$	M^{PHO}	$M^{PHO\text{-}i}$	inter SIDIS-CJCS	intra SIDIS	intra CJCS	TOTAL
1			X		X		224.200	329	/	224.529
2		X	X		X		44.567	0	/	44.567
3			X				3.066	78	4.990	8.134
4	X		X		X		8.110	17	/	8.127
5	X	X	X		X		7.106	0	/	7.106
6	X	X	X				23	0	6.374	6.397
7		X	X				176	0	6.012	6.188
8	X		X				56	5	2.196	2.257
9				X		X	991	96	618	1.705
10					X		1.063	52	/	1.115
11		X					254	26	268	548
12	X						65	164	296	525
13				X			44	58	394	496
14						X	32	54	362	448
15			X	X	X	X	372	3	/	375
16	X	X					29	2	164	195
17		X			X		107	0	/	107

Un examen du Tableau 23 mène aux constats suivants. Tout d'abord, la majorité des liaisons (environ 9 liaisons sur 10) sont établies entre SIDIS et CJCS.

Dans cette orientation inter-SIDIS-CJCS, 9 situations sur 10 sont de type $[M^{TRI}+M^{PHO}]$ ou $[M^{RRN}+M^{TRI}+M^{PHO}]$.

On réalise que les méthodes M^{TRI} et M^{PHO} sont omniprésentes. Par ailleurs, dans l'orientation inter-SIDIS-CJCS, elles apparaissent quasi toujours simultanément. On a toutefois quelques milliers de cas où

[42] Notez que pour les liens Intra-CJCS, il n'y a pas de chiffres pour les situations impliquant la méthode M^{PHO}, car elle n'a pas été appliquée dans ce cas de figure. On indique cette absence d'information par une barre "/" et non pas par 0.

la méthode M^{TRI} apparait seule, et où M^{PHO} apparait seule, signe que des méthodes ne sont pas strictement équivalentes.

La méthode M^{RRN} est généralement accompagné par M^{TRI} ou M^{PHO}. Les cas où elle apparaît seule – [M^{RRN}] – sont rares (< 1%). Pareillement, la méthode M^{JUG} apparaît en accompagnement des autres méthodes, et les cas où elle apparaît seule – [M^{JUG}] – sont du même ordre de rareté que le scénario [M^{RRN}].

Enfin, il y a quelques rares cas de méthodes inversées (M^{TRI-i} et M^{PHO-i}) qui ne sont en majorité pas accompagnées d'autres méthodes.

En résumé, toutes les méthodes sont capables (dans le sens de leur capacité à établir des liens), mais elles ne sont pas équivalentes. Toutes sont capables de récupérer des liens à elles toutes seules (i.e., sans être accompagnées d'autres méthodes) et donc aucune n'est redondante dans le sens absolu du terme. Toutefois, et c'est sans doute le plus important, certaines sont capables plus que d'autres de récupérer un grand nombre de liens, telles que les méthodes M^{TRI} et M^{PHO}.

Dans un travail ultérieur, il pourrait être intéressant d'analyser la qualité des liaisons obtenues en inspectant un échantillon de liens dans chacune des 46 situations qui se sont manifestées. Une attention particulière pourra être donnée aux situations où une méthode se manifeste isolément, afin de vérifier que le lien qui est récupéré est bien crédible et non une anomalie.

3.4. Étape 4 : les nœuds de personnes

Un nœud « personne » peut se trouver associé à une seule source d'enregistrements (soit SIDIS, soit CJCS) ou aux deux sources d'enregistrements en même temps (i.e., SIDIS et CJCS). Nous avons donc trois possibilités : SIDIS et CJCS ensemble, SIDIS seul, et CJCS seul.

Nous avons compté le nombre de personnes pour chaque possibilité dans le cadre des 24 scénarios (voir Tableau 59, Tableau 61, Tableau 63, et Tableau 65 en annexe), et en avons tirés différents constats et différentes figures, que nous présentons ci-après.

3.4.1. Les personnes qui sont dans SIDIS et CJCS

La situation qui nous intéresse le plus est celle où une personne à au moins un enregistrement dans SIDIS et au moins un enregistrement dans CJCS. En fait, c'est grâce à ce cas de figure qu'il sera possible de reconstruire la carrière criminelle d'une personne en prenant en compte à la fois des données de détention (SIDIS) et des données de condamnation (CJCS).

Le nombre de personnes qui sont à la fois dans SIDIS et dans CJCS varie entre 52.711 et 296.410 (voir Figure 19). Comme on peut le voir dans la figure, ce nombre est fonction de la fenêtre de paramètres (de [L] à [RRN]), mais ne varie pas de manière perceptible selon les deux autres dimensions (exclusion ou non des liens « intra » d'une part, exclusion ou non des dossiers « inactifs » d'autre part).

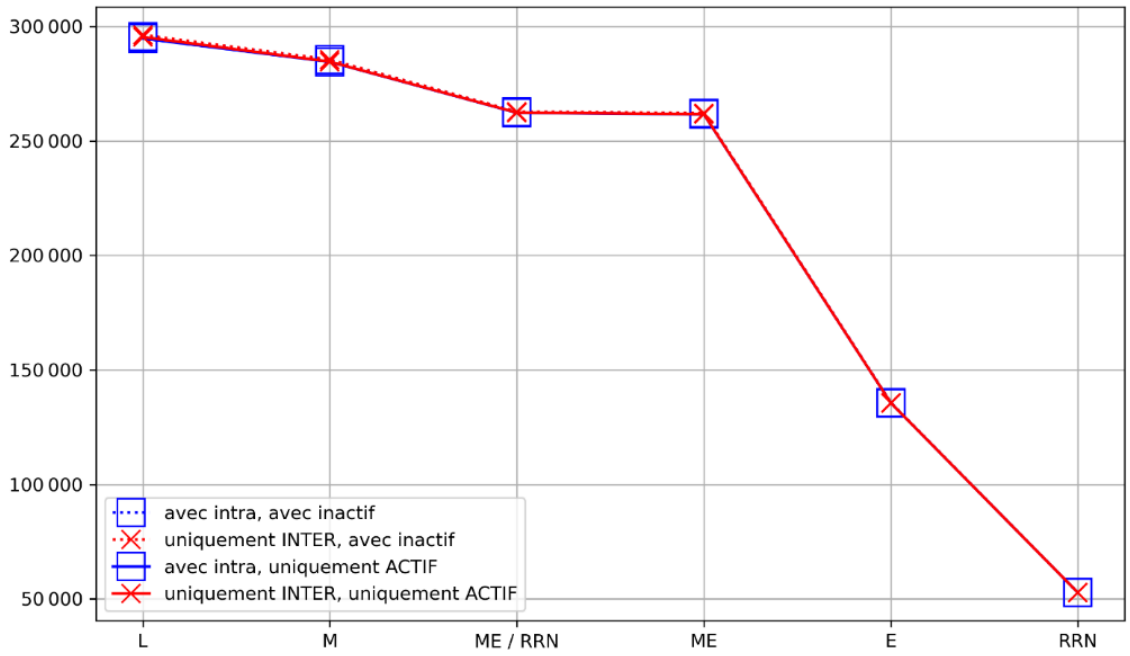


Figure 19 – Nombre de personnes à la fois dans SIDIS et CJCS (selon les vingt-quatre scénarios méthodologiques)

Dans la Figure 19, les fenêtres ont été ordonnées sur l'axe des X, de manière à être de plus en plus restrictives quand on le lit de gauche à droite, de telle façon que la fenêtre utilisée diminue de plus en plus le nombre de liens autorisés.

On commence avec la fenêtre large (L), puis la fenêtre moyenne (M), etc. La fenêtre la plus restrictive, située à droite du graphique est celle où l'on ne considère que les liens basés sur le RRN. Puisque la fenêtre devient plus étroite à mesure que l'on va vers la droite de la figure, le nombre de liens considérés entre SIDIS et CJCS diminue, et par conséquent le nombre de personnes qui sont à la fois dans SIDIS et CJCS diminue.

Comme on le voit dans la Figure 19, le nombre de personnes qu'on retrouve à la fois dans SIDIS et CJCS grâce à la méthode M^{RRN} est relativement faible (~ 50.000). Il est très vraisemblable qu'une telle méthode entraîne un grand nombre de faux négatifs.

Si l'on divise le nombre (variable) de personnes qui sont à la fois dans SIDIS et CJCS par le nombre (lui aussi variable) de personnes qui sont dans SIDIS, on obtient la proportion de personnes de SIDIS que l'on trouve également dans CJCS. Cette proportion varie entre 14,43% et 82,13% en fonction des six fenêtres considérées (voir Figure 20).

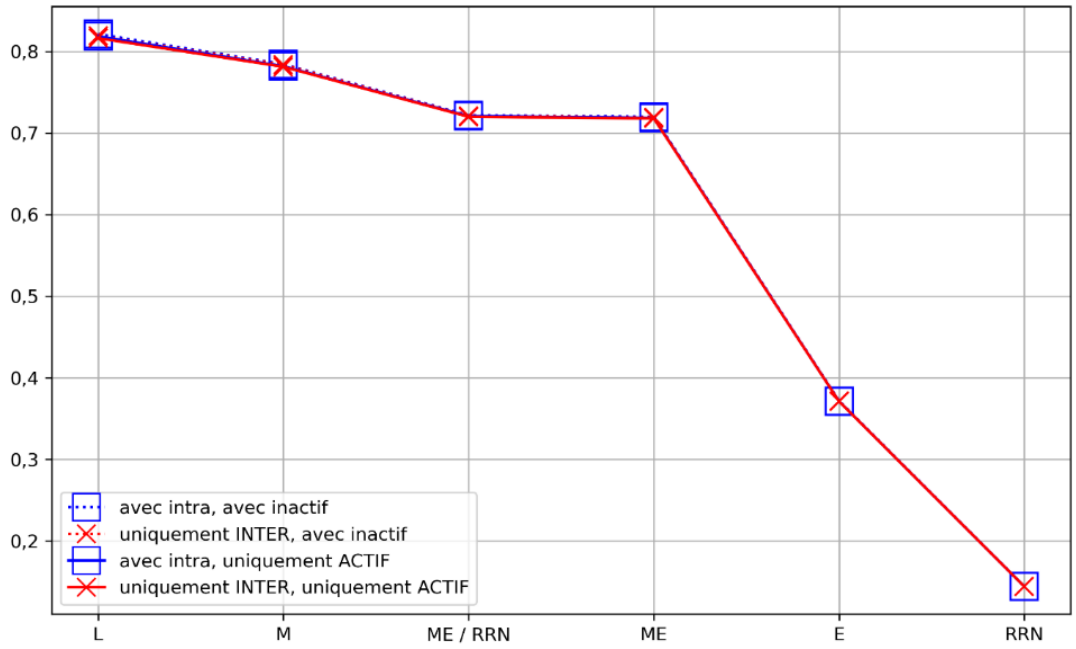


Figure 20 – Proportion des personnes de SIDIS qui sont aussi dans CJCS

Comme on peut le voir dans la Figure 20, se contenter uniquement du RRN ne permet de relier qu'environ 14% des personnes de SIDIS à CJCS. Toutefois seulement 56.220 enregistrements sur les 365.401 enregistrements existant ont un RRN, soit environ 15% des enregistrements de SIDIS.

Si l'on se concentre sur les enregistrements de SIDIS qui ont un RRN, et qu'on calcule la proportion d'entre eux qui sont liés à CJCS, on réalise qu'elle varie entre 93,78% et 93,80%, ce qui est fort proche du 93,84% rapporté par De Blander et al. (2019)[43], et démontre l'importance du RRN pour apparier les enregistrements et la nécessité de disposer de cette information ou d'une information de valeur comparable pour relier les enregistrements entre eux. Autrement dit, lorsque l'on dispose du RRN, la probabilité de pouvoir lier l'enregistrement d'une source à un celui d'une autre source est élevée (> 90%).

Si on retourne à la Figure 20 qui représente la proportion des personnes de SIDIS qui sont aussi dans CJCS, on voit que la proportion la plus élevée (82%) est atteinte avec la fenêtre [L]. Comme celle-ci contiendra vraisemblablement[44] de nombreux faux positifs, il vaut mieux l'éviter.

---

[43] Dans la section intitulée "Eerste koppeling op basis van RR-N" (premier couplage basé sur le RRN), on peut en effet lire "Een eerste koppeling op basis van het rijksregisternummer levert 52374 gematchte records op (93.84% van het aantal observaties met een rijksregisternummer in de SIDIS data)." (Un premier lien basé sur le numéro de registre national donne 52.374 enregistrements appariés (93,84% du nombre d'observations avec un numéro de registre national dans les données SIDIS).)

[44] Comme nous l'avons déjà énoncé, il serait utile de développer une méthode pour évaluer cela de manière plus objective.

La fenêtre [M], [ME] et [ME/RRN] correspondent à 78%, 72% et 72% respectivement.

La fenêtre [E] donne une proportion de 37%, soit environ la moitié de la proportion de la fenêtre [M] (78%). Si on souhaite utiliser la fenêtre [E], il faudra garder à l'esprit qu'elle entraîne vraisemblablement de nombreux faux négatifs.

Si l'on fait la supposition que la bonne fenêtre est de type moyenne (M, ME, ME/RRN), le bon chiffre se situera donc entre 72% et 78%, ce qui est plus élevé que le 58,06% rapporté par De Blander et al. (2019)[45]. Cette différence peut s'expliquer pour deux types de raisons.

Premièrement, les données utilisées par ces auteurs ne sont pas strictement similaires à celles que nous utilisons. Si les données de SIDIS qu'ils emploient sont identiques aux nôtres, en revanche les données de CJCS dont nous bénéficions sont plus récentes : les données utilisées par De Blander et al. (2019) datent de 2018 tandis que les nôtres datent de 2020. Des données plus récentes contiendront davantage de données (et sans doute aussi davantage de données corrigées) et donc globalement plus d'opportunités pour un appariement entre SIDIS et CJCS.

Deuxièmement, les techniques utilisées ne sont pas identiques. Si De Blander et al. (2019) utilisent comme nous de l'information relative aux noms et données de naissance (sexe, date de naissance, ville de naissance) ainsi que certains dérivés (e.g., obtenus via l'algorithme Double Metaphone et l'inversion des noms et prénoms), nous n'utilisons pas exactement la même information et la mécanique de l'algorithme général n'est pas identique. Par exemple, en ce qui concerne l'information utilisée, nous exploitons les dates de condamnation pour trouver des enregistrements à comparer.

Ceci étant dit, nous ne disposons pas d'information précise sur la manière dont l'information a été agencée dans la procédure d'intégration référencée dans De Blander et al. (2019), et ne pouvons donc pas nous prononcer sur l'impact comparatif des différentes techniques utilisées.

3.4.2. Les personnes qui sont uniquement dans SIDIS ou CJCS

Puisque le nombre de personnes qui sont à la fois dans SIDIS et CJCS diminue à mesure que la fenêtre devient plus restrictive (de [L] vers [RRN]), cela entraîne que le nombre de personnes qui sont uniquement dans SIDIS (voir Figure 21) ou uniquement dans CJCS (voir Figure 22), va, lui, augmenter.

[45] Dans la section intitulée "Tweede-eenenveertigste koppeling op basis van naam- en geboortegegevens" (quarante-deuxième lien basé sur le nom et les données de naissance), on peut en effet lire "Uiteindelijk levert dit 212148 gematchte records op (58.06% van het aantal observaties in de SIDIS data)" (Au final, on obtient 212.148 enregistrements appariés (58,06% du nombre d'observations dans les données SIDIS).)

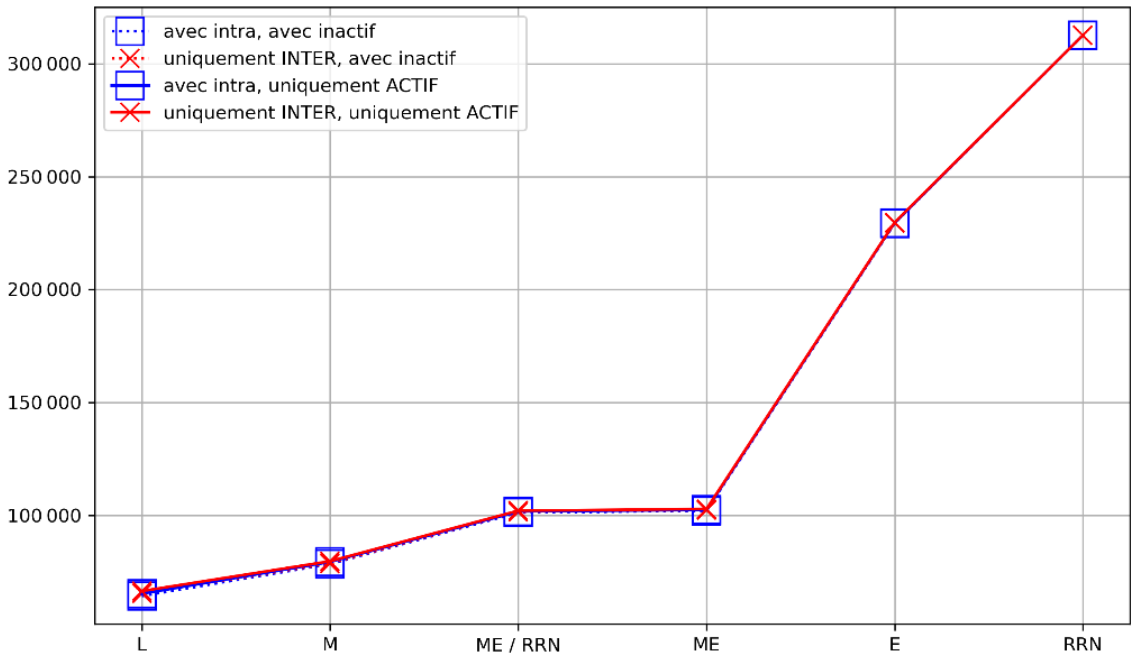


Figure 21 – Nombre de personnes qui sont uniquement dans SIDIS (selon les vingt-quatre scénarios méthodologiques)

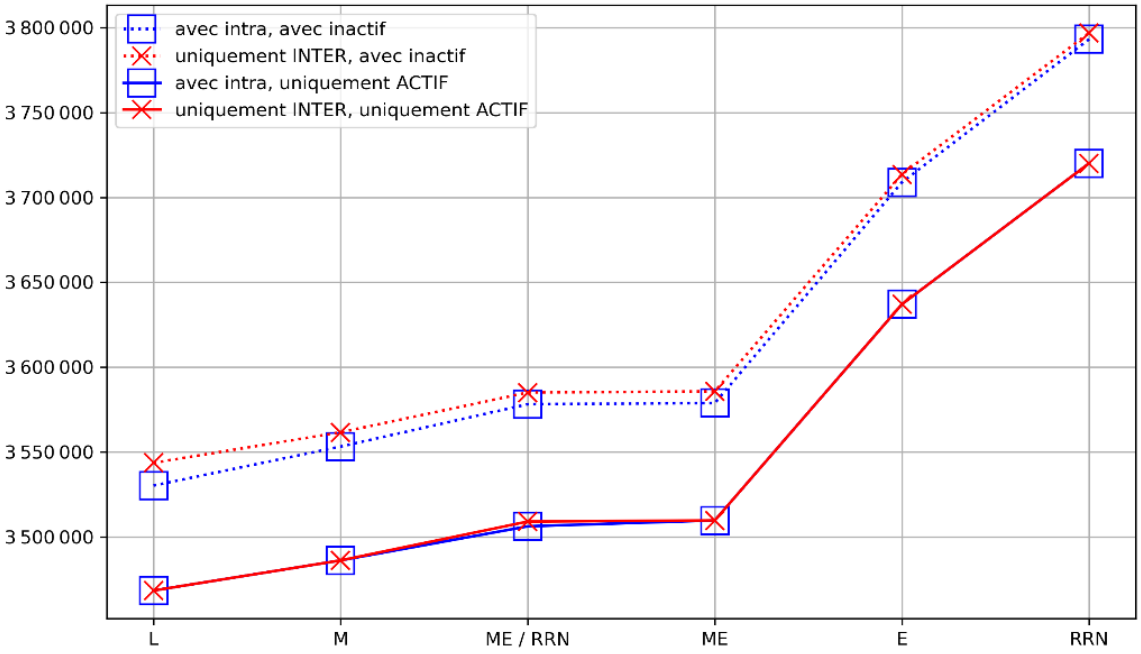


Figure 22 – Nombre de personnes qui sont uniquement dans CJCS (selon les vingt-quatre scénarios méthodologiques)

Dans la Figure 22, contrairement à ce qu'on a pu voir dans les figures précédentes, on peut observer un effet de l'exclusion des dossiers de CJCS qui ne sont pas actifs. En effet, il y a à la base 3.860.989 nœuds de dossiers dans CJCS, et parmi eux 88.212 qui ne sont pas actifs (2% du total). Quand on passe des scénarios comprenant des inactifs (lignes en pointillé en haut dans le graphique) aux scénarios

excluant les inactifs (lignes en trait plein en bas dans le graphique), on perd 2% du nombre de personnes.

En revanche, quand ignore les liens « intra » (i.e., intra-SIDIS et intra-CJCS), il y a une légère augmentation du nombre de personnes. En effet, quand on passe des scénarios comprenant des liens « intra » (lignes avec des carrés dans le graphique) aux scénarios excluant les liens « intra » (lignes avec des ronds dans le graphique), on augmente un petit peu le nombre de personnes. Cette augmentation est toutefois plus visible pour les scénarios n'excluant pas les inactifs (lignes en pointillé en haut dans le graphique) que pour les scénarios excluant les dossiers inactifs (lignes en trait plein en bas dans le graphique).

3.4.3. La condamnation définitive à une peine d'emprisonnement

On a déjà vu que la proportion des personnes de SIDIS qui sont aussi dans CJCS varie entre 14,43% et 82,13% en fonction des six fenêtres considérées (voir Figure 20).

On devrait s'attendre à ce que cette proportion augmente si l'on se concentre sur les personnes de SIDIS qui ont été condamnées définitivement à une peine d'emprisonnement. C'est effectivement ce que l'on observe puisqu'alors la proportion varie entre 18,07% et 91,39% (voir Figure 23), soit une augmentation entre 4 et 9% par rapport aux proportions de la Figure 20.

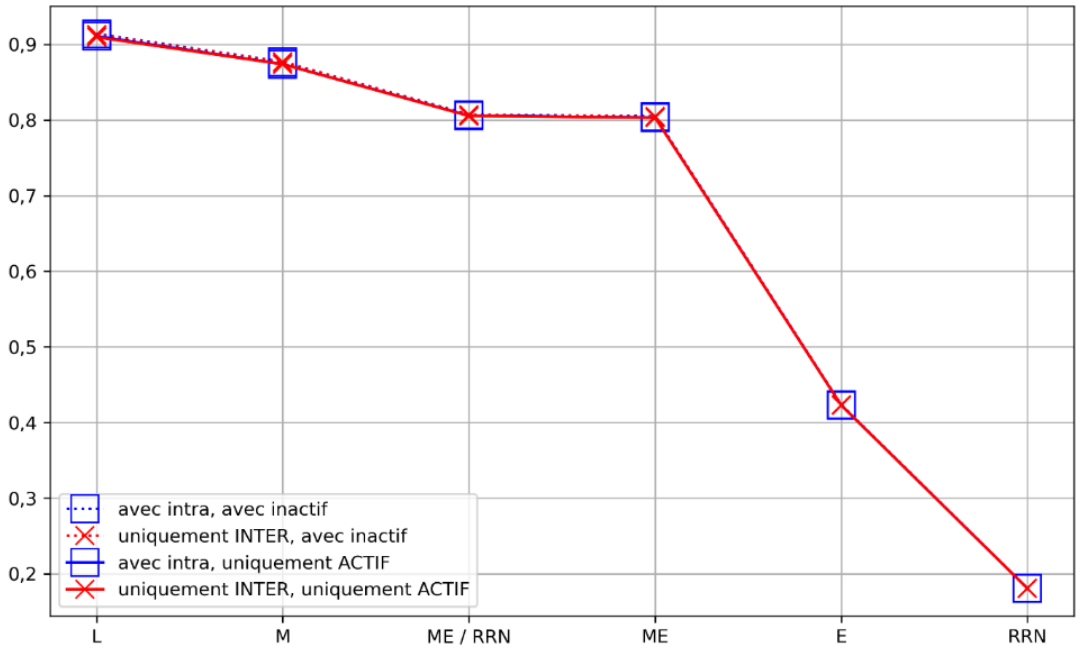


Figure 23 – Parmi les personnes de SIDIS qui ont été condamnées définitivement à une peine d'emprisonnement, proportion d'entre elles qu'on retrouve également dans CJCS

Dans l'idéal, nous nous serions attendus à ce que 100% des personnes ayant été définitivement condamnées à une peine d'emprisonnement aient au moins un enregistrement de condamnation dans CJCS. Pourquoi n'atteint-on pas 100% ? Pour le savoir, il faudra examiner les caractéristiques des personnes qui, bien qu'ayant été condamnées à une peine d'emprisonnement, n'ont pas

d'enregistrement dans CJCS (e.g., en vérifiant si ce sont des personnes décédées et en examinant les périodes de détention).

Enfin, on un certain nombre de personnes qui sont liées à CJCS et SIDIS bien qu'elles n'aient pas été condamnées à une peine d'emprisonnement. Comment expliquer ces cas ? Une hypothèse est que ces personnes n'ont effectué qu'une détention préventive (ce qui explique leur présence dans SIDIS) et ont été condamnées à une autre peine par ailleurs (ce qui explique leur présence dans CJCS). Une analyse future cherchera à élucider ces cas.

3.4.4. L'ambiguïté des enregistrements

Un nœud « personne » peut avoir plusieurs enregistrements dans une même base de données (i.e., intra-SIDIS ou intra-CJCS). Par exemple une personne peut avoir deux enregistrements dans SIDIS, tandis qu'une autre peut avoir trois enregistrements dans CJCS. Si une personne a plusieurs enregistrements dans SIDIS, on a affaire à un cas « ambigu » (c'est-à-dire potentiellement problématique). Si une personne a plusieurs enregistrements dans CJCS, on a également affaire à un cas ambigu, sauf si seul l'un de ces enregistrements est actif (e.g., on a pour la même personne un dossier actif et deux dossiers effacés).

Il faut noter que d'un point de vue technique, on pourra avoir des cas ambigus, même lorsqu'on ne trace pas de lien « intra » mais qu'on se contente uniquement de tracer des liens inter-SIDIS-CJCS. En effet, on pourrait par exemple avoir une personne qui possède un enregistrement dans CJCS et deux enregistrements dans SIDIS, par l'entremise des liens inter-SIDIS-CJCS.

Il y a deux raisons pour lesquelles on peut avoir une situation d'ambiguïté dans les enregistrements d'une personne. Premièrement, il peut s'agir de duplicatas véritables, qui n'ont pas été détectés par le gestionnaire de la base de données source (e.g., SIDIS) mais qui ont été détectés via l'analyse réalisée dans l'IHD. On a ici bien affaire à la même personne. Deuxièmement, il peut s'agir d'enregistrements inclus de manière erronée par l'IHD et dans ce cas on n'a pas affaire à la même personne (i.e., on fait face à des faux positifs).

Les cas ambigus sont potentiellement problématiques car il n'est pas facile de décider de manière certaine qu'on a affaire ou non à des personnes différentes. Combien de personnes présentent des configurations ambiguës de ce genre ?

La grande majorité des personnes (> 99%) ne correspondent pas à des cas ambigus. En effet, parmi les scénarios considérés au plus 18.768 personnes sont concernées par ce phénomène[46]. Comme on le voit dans la Figure 24, c'est la fenêtre la plus permissive, la fenêtre large [L], qui correspond à ce maximum. Il est vraisemblable qu'un certain nombre d'entre eux correspondent à des faux positifs.

[46] L'ensemble des chiffres est disponible dans le Tableau 58 de l'annexe dans la section A.7.

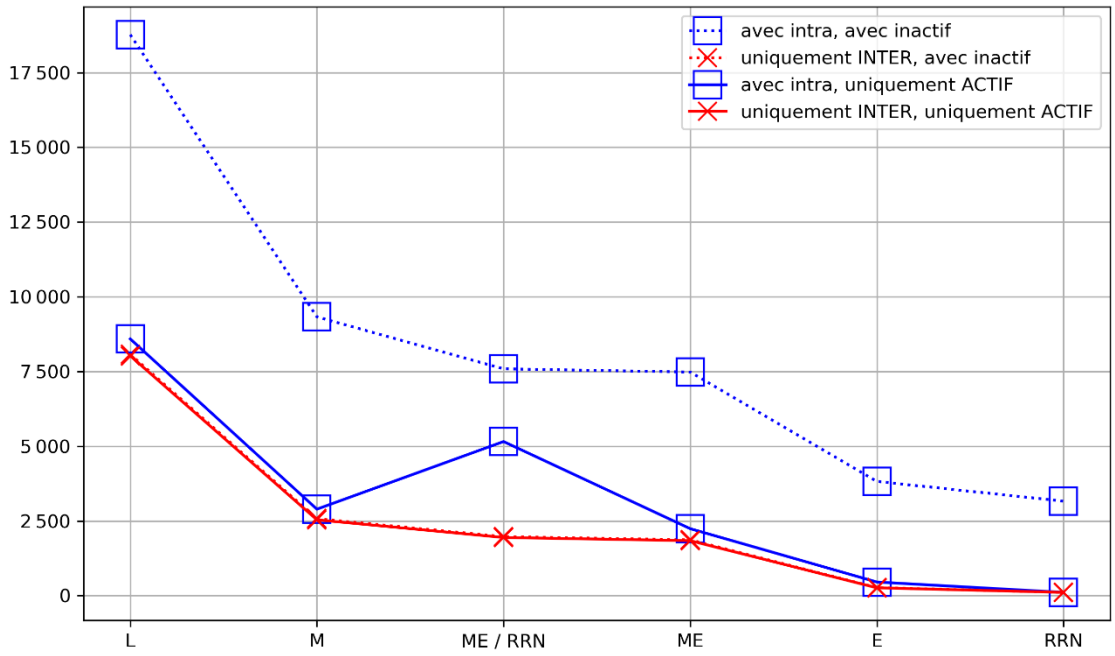


Figure 24 – Nombre de personnes correspondant à des cas ambigus

Ensuite, les scénarios qui prennent en compte les liens « intra » et les dossiers « inactifs », voient une augmentation de 3.000 à 10.000 dans le nombre de personnes dont les enregistrements sont ambigus (cf. ligne en pointillé en haut du graphique).

Enfin, dans le scénario où l'on inclut tous les liens de type RRN en plus des liens de qualité moyenne (scénario [ME/RRN]), le fait de prendre les liens « intra » en considération double le nombre de cas ambigus (on passe d'environ 2500 à 5000 personnes aux enregistrements ambigus).

Les raisons qui mènent à ces cas ambigus doivent encore être investiguées. En attendant que lumière soit faite, ces quelques personnes pourront être exclues ou faire l'objet d'un traitement séparé dans les études qui seront conduites dans le futur.

4. Conclusion et discussion

Le présent rapport visait à rendre compte du développement de la base de données historique intégrée – ou Integrated Historical Database (IHD) – un entrepôt de données et outil permettant de préserver, intégrer et exploiter les données extraites de deux bases de données relatives à l'administration de la justice pénale belge.

4.1. Les données sources

La première extraction obtenue en octobre 2020 concerne les décisions judiciaires définitives prononcées à l'encontre d'environ 3 millions de personnes telles qu'elles ont été enregistrées dans le Casier Judiciaire Central (CJCS). Le deuxième jeu de données, obtenu en 2014 concerne les informations relatives à la détention d'environ 300.000 personnes, telles qu'elles ont été enregistrées dans le Système d'information des détentions – greffe (SIDIS-greffe) de la Direction Générale des Établissements Pénitentiaires (DG EPI).

Ce sont des données historiques à plus d'un titre. Tout d'abord, elles concernent des périodes relativement longues des activités de l'administration pénale belge. Les condamnations répertoriées dans les données obtenues de CJCS courent de 1922 à 2020, tandis que les détentions répertoriées dans les données de SIDIS vont de 1974 à 2014. Ensuite, ces données sont arrêtées dans le temps puisqu'il s'agit d'extractions de ces systèmes d'enregistrement. Enfin, ce sont des données sur des personnes – leur trajectoire et leur histoire – bien que nécessairement partielles, car reflétant avant tout le fonctionnement de la justice pénale dont elles ont fait l'expérience.

La base de données SIDIS-greffe n'est plus opérationnelle depuis 2014, car elle a été remplacée par une nouvelle base de données : SIDIS-suite. Lorsque la migration a eu lieu, les données de SIDIS-greffe ont été importées dans SIDIS-suite, puis SIDIS-greffe a été figée. La sauvegarde au sein de l'IHD de la totalité des données de SIDIS-greffe, telles qu'elles ont été transmises par la DG EPI en 2014, est donc le témoin d'une époque car si vraisemblablement la majorité des données ont survécu lors de leur passage à SIDIS-suite, il n'en est pas de même de la structure de la base de données.

4.2. La problématique de la préservation et exploitation des données

Le fait de disposer d'extraction de données sous la forme de fichiers séparés, stockés en lieu sûr, préserve certes les données de départ, mais n'en facilite pas l'exploitation. Encore faut-il pouvoir en effet préparer ces données en vue de l'analyse. Il faut notamment utiliser la connaissance relative à la structure des données afin que les différents enregistrements contenus dans ces fichiers soient liés correctement les uns aux autres. Par ailleurs des connaissances supplémentaires peuvent venir éclairer utilement les analyses par rapport aux données utilisées (e.g., par rapport aux codes de faits dans CJCS). C'est cet ensemble de données et de connaissances que l'IHD se propose de préserver et rendre plus aisément disponibles au sein d'un même système.

4.3. La problématique de l'intégration des données

Par ailleurs, il ne s'agissait pas d'exploiter les données de CJCS et SIDIS de manière indépendante l'une de l'autre, mais de les exploiter de manière conjointe, i.e., de manière intégrée.

La problématique de l'articulation des bases de données de la justice pénale en vue de leur exploitation conjointe ne date pas d'hier (Mine & Vanneste, 2011). Un constat majeur de ce rapport est qu'en Belgique une telle articulation est rendue difficile par l'absence d'un identifiant unique et global de la personne, que ce soit au sein de la justice ou aux autres niveaux de l'État. À cause de cela, il est nécessaire de se reposer sur d'autres informations personnelles telles que le prénom, le nom et la date de naissance de la personne. Or une telle approche ne garantit pas de pouvoir faire les rapprochements corrects entre enregistrements relatifs aux personnes, car ces informations peuvent être encodées de manière variée, d'un enregistrement à un autre, et d'une base de données à une autre. Par exemple, l'information peut être incomplète dans un enregistrement, ou orthographiée différemment dans un autre enregistrement. Par ailleurs certaines personnes ont les mêmes prénom, nom de famille et date de naissance. D'autres informations sont donc nécessaires pour les différencier.

Si cette situation pose des problèmes pour produire une statistique intégrée ou mener des recherches nécessitant une intégration des données, elle est aussi problématique par rapport à des besoins plus opérationnels.

L'actualité récente s'est d'ailleurs fait l'écho de cette situation. Le 16 octobre 2023 deux supporters de football suédois étaient abattus par Abdessalem Lassoued. Évadé d'une prison tunisienne en 2011, il faisait pourtant l'objet d'une demande d'extradition vers la Tunisie, ce qui laisse à penser que l'attentat aurait pu être évité si cet homme avait pu être expulsé de Belgique. Or le rapport présenté par le Comité P au sujet de cette affaire en avril 2024 a révélé qu'un problème d'orthographe dans l'enregistrement de son nom dans les différentes bases de données de l'État a joué un rôle dans l'incapacité de ses services à répondre favorablement à cette demande d'extradition :

« L'un des problèmes qui se posent est celui de l'orthographe du nom du terroriste, écrit au début Abdesslem Laswad, puis Abdes(s)alem Lassoued. Parmi les nombreuses bases de données disponibles, certaines permettent une identification phonétique, d'autres pas, d'autres encore de façon optionnelle »[47].

Comme on l'a montré dans le présent rapport, l'utilisation d'une méthode d'identification phonétique est très utile pour relier des enregistrements entre eux sur la base du nom et du prénom. Pour prendre cet exemple ayant tragiquement fait l'actualité, « Abdesslem » et « Abdessalem » sont tous deux représentés par « APTS » par l'algorithme phonétique que nous avons employé (double Metaphone), tandis que « Laswad » et « Lassoued » sont, eux, représentés par « LST », rendant ainsi un rapprochement possible entre ces différentes variantes. Encore faut-il pouvoir appliquer cette technologie, partout où elle est nécessaire.

[47] https://www.rtl.be/actu/belgique/societe/le-comite-p-presente-son-rapport-la-chambre-concernant-lattentat-du-16-octobre/2024-04-04/article/655031

Il faut aussi pouvoir en faire usage de manière appropriée. Il y a en effet des cas de figure où la technologie se montrera impuissante pour atteindre les objectifs escomptés. Pour rester dans le contexte de cette affaire, la méthode de la représentation phonétique fonctionnera dans la mesure où les variantes orthographiques d'un mot correspondent bien à la même représentation phonétique (e.g., « Philippe » et « Filip » qui correspondent tous les deux à « FLP »). Or ce n'est pas toujours le cas (e.g., « Ahmed » correspond à « AMT » mais « Hamed » correspond à « HMT » : un déplacement minime de la lettre 'h' est fatal). Pour contourner ces obstacles d'autres algorithmes doivent être utilisés, en complément ou en substitution des précédents (e.g., via d'autres mesures de similarité textuelle).

Pour prendre un dernier exemple, rencontré dans le cadre de cette étude, dans certains cas, quoique rares, les noms et prénoms peuvent être intervertis dans les données, rendant de tels rapprochements phonétiques caducs. De manière générale, un encodage de qualité, passant par une méthodologie de traitement de l'information appropriée, ainsi qu'une politique de vérification de la qualité des enregistrements, sont indispensables.

Bien sûr, ces mesures seront d'autant moins nécessaires qu'un identifiant unique et global de la personne s'imposera dans les systèmes d'enregistrement. Si le numéro de registre national n'est pas appelé à s'étendre et à jouer ce rôle, d'autres identifiants pourraient servir à cet usage. On pense par exemple à l'identifiant résultant de l'utilisation de l'APFIS[48].

4.4. Une base de données historique intégrée

En termes d'exploitation des données, l'IHD permet d'explorer aisément les données qu'elle contient et de réaliser des calculs statistiques pour différents besoins, parmi lesquels l'analyse de la récidive et des carrières criminelles. Pour rendre cela possible, l'IHD a été développée de manière à pouvoir intégrer au niveau de la personne ces deux jeux de données provenant de sources distinctes. Cela signifie qu'elle permet de déterminer quelles personnes ont quel enregistrement dans quelle source de données (SIDIS, CJCS ou les deux). Une telle intégration permet de raisonner en termes de personne et de suivre son parcours : soit son parcours en termes de détention (dans SIDIS), soit son parcours en termes de condamnation (dans CJCS), soit son parcours en combinant de l'information sur les détentions et les condamnations (i.e., SIDIS et CJCS ensemble).

Nous abordons à présent les aspects méthodologiques les plus intéressants de l'IHD ainsi que quelques résultats issus de son utilisation.

L'IHD est tout d'abord une base de données qui permet d'exploiter les enregistrements de jeux de données provenant originellement de sources distinctes. En ce sens, elle permet de stocker les données et donc de les préserver sous un autre format que celui des fichiers d'extraction source et d'y ajouter des considérations additionnelles (i.e., structure des données, données contextuelles).

Elle utilise aussi la technologie de la base de données en graphe Neo4j. Dans une base de données en graphe, les enregistrements sont représentés par des objets – les « nœuds » – reliés entre eux par des

[48] Automated Palm and Fingerprints Identification System.

relations. Un nœud peut représenter n'importe quel concept d'intérêt. Par exemple un enregistrement de personne condamnée, un enregistrement de personne détenue, une personne. L'ensemble des nœuds et des relations est un graphe.

Le graphe est utilisé à tout moment. Les données de départ forment un graphe. Puis des nœuds et relations additionnels viennent enrichir ce graphe initial pour atteindre l'objectif final de l'intégration des données. Tout d'abord les enregistrements de personnes dans SIDIS et CJCS sont associés à des nouveaux nœuds, dits nœuds d'intégration, qui contiennent une version standardisée d'informations susceptibles d'aider à reconnaître les personnes en vue de l'intégration : information sur le nom, le prénom, la date de naissance, le numéro de registre national, etc. Ensuite, des enregistrements sont choisis pour être comparés grâce à ces nœuds d'intégration. Quand deux enregistrements sont jugés comme étant suffisamment similaires, un lien d'intégration est tissé entre eux. Enfin, des nœuds de personnes sont créés sur la base de ces liens d'intégration.

Dans ce travail, nous avons bénéficié de la flexibilité de la représentation de l'information dans Neo4j, notamment afin de tester différentes techniques d'intégration. Nous avons également utilisé la capacité de Neo4j à naviguer rapidement dans le graphe pour récupérer une grande quantité d'enregistrements à évaluer, mais aussi limiter cette quantité à un volume raisonnable.

Six méthodes ont été mises au point pour trouver des enregistrements des personnes et les comparer. Globalement chaque méthode présente des qualités dans la mesure où chacune sera capable de rapprocher des enregistrements que d'autres méthodes ne parviendront pas à rapprocher. Cependant, certaines méthodes permettent de créer beaucoup plus de liens que d'autres. Leur usage est indispensable.

Deux enregistrements de personnes peuvent être liés par un ou plusieurs liens d'intégrations. En général, les liens sont de qualité variable, car la similarité des enregistrements et la quantité de preuves (i.e., informations comparées : nom, prénom, date de naissance, etc.) pour juger de leur similarité, varient. Une procédure déterministe a été développée pour donner un certain poids aux preuves, et calculer des scores de similarité basés sur ces preuves. Ensuite, puisque les liens présentent une qualité variable, les liens réalisés ont été examinés selon différents critères de qualité. Par ailleurs, 24 scénarios méthodologiques ont été élaborés pour choisir quels enregistrements et quels liens entre enregistrements permettent de distinguer des personnes. Ces scénarios permettent d'étudier l'impact des choix méthodologiques et rester prudent quant à l'interprétation des résultats.

En attendant, selon les scénarios retenus, entre 1% et 8% des données, soit entre 50.000 et 300.000 personnes, ont été identifiées, qui présentent à la fois des données de détention (dans SIDIS) et des données de condamnation (dans CJCS). Au terme de son développement, l'IHD est donc en mesure d'offrir des jeux de données centrés sur un nombre non négligeable de personnes aux fins de la recherche scientifique sur la récidive et les carrières criminelles.

En effet, par rapport à l'exercice d'intégration précédent, rapporté par De Blander et al. (2019), l'IHD utilise une méthodologie d'intégration des données qui est transparente et réversible. Dans l'exercice précédent, l'intégration des données était relativement peu documentée et avait donné lieu à un unique fichier de données où les informations pertinentes de SIDIS et CJCS se trouvaient fusionnées selon une méthode de fusion particulière. Il n'était dès lors plus possible de questionner les choix opérés pour réaliser l'intégration, ceux-ci n'étant pas toujours clairs, ni de revenir en arrière en

appliquant d'autres choix, l'unique procédure d'intégration suivie étant contenue dans des scripts dépourvus de manuel d'utilisation.

Contrairement à l'exercice précédent, l'IHD n'est pas un fichier de données mais une base de données proprement dite, dont la construction est documentée, et où apparaissent clairement les données originales et les éléments additionnels au travers desquels se manifestent différents scénarios d'intégration des données (i.e., nœuds d'intégration, liens d'intégration, et nœuds de personne). Il est donc possible d'en comprendre le fonctionnement, de tester différents cas de figure et de revenir en arrière si le besoin s'en faisait sentir.

4.5. Limites actuelles et développements futurs

À l'heure actuelle, de multiples interrogations demeurent, qui sont autant d'aspects qui pourront être abordés dans le futur, en nous aidant de l'IHD et de tout développement ultérieur : ceux de la Cellule Récidive et Carrières Criminelles (https://incc.fgov.be/CReCC), en général, et, en particulier, ceux de la Database on Offender Trajectories (DOT), qui est en cours de développement dans le projet BELSPO du même nom (https://incc.fgov.be/DOT). Nous évoquons ces interrogations ci-dessous.

1) Nous n'avons pas encore une vue précise sur les données de SIDIS-greffe qui sont passées dans SIDIS-suite et quelles innovations en termes de structure de base de données ont eu lieu lors du passage à SIDIS-suite. Il s'agira d'examinera ce passage, en incluant une extraction de SIDIS-suite. Cette prise en compte des données de SIDIS-suite sera l'occasion de compléter les données de détention afin de pouvoir lier davantage d'enregistrements des personnes qui ont fait de la détention à leurs enregistrements dans CJCS. En effet, les données de SIDIS-greffe s'arrêtent en 2014, ce qui pose un problème de rupture dans la série temporelle (e.g., certaines personnes qui ont des enregistrements dans CJCS devraient avoir des données de détention, mais celles-ci ne sont pas disponibles car les données de SIDIS-greffe ne contiennent pas ces enregistrements : pour trouver ces enregistrements, les données de SIDIS-suite sont nécessaires).

2) Pareillement, la prise en compte d'une nouvelle extraction de données de CJCS, permettra de bénéficier des dernières innovations en matière de gestion des données de condamnation. On pense par exemple au travail effectué par le Casier pour clarifier le statut « effacé » de certains éléments du jugement (voir Huynen, Mine et al, 2024).

3) L'IHD intègre les données via une méthode déterministe établie via une série de choix méthodologiques. Il serait bienvenu de mettre au point une méthode probabiliste qui exploite les données de manière optimale. Par ailleurs, l'intégration des données a jusqu'ici essentiellement exploité des données à caractère personnel (nom, prénom, RRN, …) pour établir des rapprochements entre enregistrements. Des données propres aux carrières des personnes (en termes de condamnation et d'emprisonnement), autres que la simple information sur une date de jugement, pourraient être utilisées pour améliorer la qualité de ces rapprochements.

4) Dans le cadre du développement de l'IHD, tous les enregistrements ont été stockés comme s'ils étaient composés de champs de type textuel. Cela a permis de pouvoir développer le système, avant que ne soit disponible l'information complète sur les types des champs (e.g., nombre entier, date) et sans que cela n'ait aucun impact sur la procédure d'intégration. Toutefois, un tel choix fait que les données prennent plus de place que nécessaire sur le disque. Dans une prochaine version du système,

les types appropriés seront ajoutés aux données. On notera que cet ajout peut se faire sans difficulté en aval du processus de création de l'IHD, puisque Neo4j stocke les propriétés des nœuds de manière indépendante et qu'il n'existe pas de définition générale des types qui s'applique à l'ensemble des données.

5) Un examen des dates de jugement doit être conduit afin de comprendre pourquoi cette information est en apparence si complète au niveau de SIDIS. Il sera par ailleurs utile de déterminer dans CJCS la date de jugement sous un format plus précis que le format dd-MMM-yy (e.g., 18-JAN-02), où l'année ne comporte que les deux derniers chiffres.

6) Un travail de standardisation des adresses des lieux de résidence pourra être fait afin de récupérer les coordonnées géographiques, et faciliter l'intégration des données.

7) Un nouveau travail de standardisation pourra aussi être effectué avec le lieu de naissance et le lieu de résidence de SIDIS. En effet, pour l'instant, la standardisation est sommaire puisque CJCS dispose de la traduction en quatre langues (français, néerlandais, allemand, anglais). Un lien peut donc être fait aisément entre l'une de ces traductions et le mot encodé dans SIDIS. Toutefois, avec l'adjonction d'une autre base de données, un tel lien avec SIDIS pourrait ne pas pouvoir être effectué. Un travail de standardisation d'application plus générale sera alors nécessaire.

8) Nous avons vu qu'il existait plusieurs milliers de cas de personnes ambiguës ayant plusieurs enregistrements actifs dans une même base de données. De tels cas doivent être analysés afin d'en comprendre la portée.

9) Aucun contrôle de la qualité formelle des RRN n'a été effectué dans le présent exercice. Une telle opération permettrait peut-être de traiter différemment les nœuds associés à des RRN de mauvaise qualité (e.g., qui ont été mal encodés). Si un RRN formellement incorrect a été fautivement encodé dans un enregistrement et qu'il est par hasard le même qu'un RRN correctement attribué à un autre enregistrement, il y a un risque de faux positif (même si le risque est faible dans la mesure où les autres informations ne correspondront sans doute pas). Par ailleurs, aucune différence n'est faite entre les RRN associés aux dossiers actifs et inactifs, ou qui appartiennent à l'historique de la gestion du dossier. De telles considérations pourraient éventuellement nourrir la procédure d'intégration.

10) Il serait bon d'exploiter l'analyse des enregistrements ambigus (personnes associées à plusieurs enregistrements), et tout spécialement les personnes associées à un grand nombre d'enregistrements pour détecter d'éventuelles valeurs aberrantes. Une telle analyse a révélé jusqu'ici que certaines valeurs types exprimaient des données manquantes (e.g., « onbekend » dans les champs de nom et prénom). La présence de valeurs de ce genre risque en effet de grouper ensemble des enregistrements qui ne devraient pas l'être (e.g., des enregistrements de personnes appelées « onbekend » qui sont groupés ensemble alors qu'ils appartiennent en réalité à des personnes différentes).

11) Dans notre approche des noms et prénoms, les prénoms ou noms constitués d'un très petit nombre de lettres (par exemple plus petits que 3), pourraient devoir recevoir une attention particulière. Peut-être s'agit-il dans certains cas d'anomalies que l'on souhaite exclure. Par ailleurs, les prénoms composés et les noms composés ont été traités comme des entités distinctes. Par exemple si un champ prénom contenait le mot « Jean Christophe » (sans tiret), deux prénoms ont été associés à l'enregistrement « Jean » et « Christophe ». Même si on peut lier « Jean » et « Christophe » aux

prénoms de tout autre enregistrement, on perd malgré tout la spécificité du prénom « Jean Christophe », dans cet ordre particulier.

12) Dans cette étude, l'accent a été mis sur l'exploration de différentes méthodes et scénarios pour mesurer l'impact des choix réalisés sur les chiffres d'intérêt. Il y a pourtant différentes sections du programme qui pourraient être améliorées pour augmenter la vitesse d'exécution, notamment en plaçant chaque enregistrement qui a été examiné une fois en mémoire vive, de façon à ne pas devoir le récupérer systématiquement dans la base de données pour chaque comparaison future le concernant.

13) Nous n'avons pas exploré de manière exhaustive l'espace des paramètres des poids et seuils, ni dans l'étape 3 pour créer des liens, ni dans l'étape 4 pour créer des nœuds de personnes. Il serait le bienvenu de mettre au point une procédure qui investigue de manière plus systématique l'impact de ces paramètres. En particulier, il sera utile de définir ce qui constitue un bon lien, par exemple en termes qualitatifs (e.g., « il faut que le prénom, et le nom soient suffisamment semblables et que … »), ou du moins définir l'impact des paramètres actuels par rapport à des considérations qualitatives. Un examen de cet impact pourra se faire dans différentes régions des valeurs des paramètres : des régions très sures (e.g., proche de 90% de similarité) et des régions moins sures (e.g., proche de 70% de similarité). On pourra également examiner la qualité des liens en fonction des différentes situations de liaisons observées (46 situations sur les 63 situations possibles), et notamment ces situations où une méthode se manifeste isolément (e.g., $M^{TRI\text{-}i}$ ou $M^{PHO\text{-}i}$). On pourra ainsi éventuellement détecter des anomalies dans les liens formés.

14) Il est clair qu'il n'y a in fine pas d'intérêt à chercher à relier des personnes morales à des données de détention. Il serait facile d'empêcher de former des liens entre de tels enregistrements en prévoyant explicitement ce cas de figure à éviter dans la procédure d'intégration.

15) Dans le présent exercice, nous n'avons pas mené jusqu'à son terme, via la méthode M^{PHO}, l'exploration en phase de dégrossissage des liens intra-CJCS. Il pourrait être utile de relancer la procédure en la révisant pour la rendre plus efficace, en la faisant tourner sur une machine plus puissante ou en l'exécutant sur une plus longue période. Cela permettra ainsi de compléter nos données, afin de comparer au mieux les six méthodes l'une à l'autre.

16) Dans notre examen du nombre de personnes de SIDIS qui sont liées à CJCS, nous ne sommes pas parvenus à 100% de connexion à CJCS. Nous n'y sommes même pas arrivés en nous concentrant uniquement sur ces personnes de SIDIS qui avaient été définitivement condamnées à une peine d'emprisonnement. A contrario, on observe des personnes de SIDIS qui sont liées à CJCS, bien qu'elles n'aient pas été condamnées à une peine d'emprisonnement. Il nous faudra étudier ces cas de figure.

17) Nous n'avons pas développé de méthode pour mesurer le nombre de faux positifs et le nombre de faux négatifs. C'est un écueil auquel il faudrait chercher à remédier, à imaginer que cela soit possible en l'absence de données extérieures pour affirmer que des enregistrements sont bien correctement liés les uns aux autres.

18) Le fait que nous ayons défini 24 scénarios méthodologiques pour définir des nœuds de personnes, est un appel à la prudence par rapport à l'examen des résultats. C'est aussi un dispositif qui permet d'explorer ces résultats afin de mieux comprendre l'impact des choix méthodologiques effectués. Cela

ne signifie en rien que d'autres choix méthodologiques, meilleurs que ceux retenus dans le présent exercice, ne pourraient pas être faits.

19) Ce rapport contient un ensemble de chiffres relativement restreint par rapport à la question de recherche relative aux caractéristiques des personnes. Quelle est la part des hommes et des femmes ? Quelles nationalités sont impliquées ? Quelles sont les infractions dans lesquelles elles sont impliquées ? Nous répondrons à toutes ces questions dans des travaux ultérieurs.

20) Enfin, nous étant focalisé prioritairement sur le nombre de personnes que nous étions capables de relier de SIDIS à CJCS, nous avons postposé la résolution des questions qui pourtant nous intéressent le plus : dans quelle mesure récidivent-elles ? À quoi ressemble leur carrière criminelle ? Tant de questions auxquelles nous finirons par répondre avec les données dont nous disposons et les méthodes que nous mettrons au point.

5. Références

Blumstein, Alfred, Jacqueline Cohen, Jeffrey A. Roth, et Christy A. Visher, éd. 1986. *Criminal Careers and « Career Criminals »: Volume I. Panel on Research on Criminal Careers*. Washington, DC: The National Academies Press.

Cheng, Yijian, Pengjie Ding, Tongtong Wang, Wei Lu, et Xiaoyong Du. 2019. « Which Category Is Better: Benchmarking Relational and Graph Database Management Systems ». *Data Science and Engineering* 4(4):309-22. doi: 10.1007/s41019-019-00110-3.

De Blander, Rembert, Luc Robert, Christophe Mincke, Eric Maes, et Benjamin Mine. 2019. *Étude de faisabilité d'un moniteur de la récidive/Haalbaarheidsstudie betreffende een recidivemonitor*. *Rapport de recherche/Eindrapport*. 42. Brussels: NICC.

Dunn, Halbert L. 1946. « Record Linkage ». *American Journal of Public Health and the Nations Health* 36(12):1412-16. doi: 10.2105/AJPH.36.12.1412.

Fellegi, Ivan P., et Alan B. Sunter. 1969. « A Theory for Record Linkage ». *Journal of the American Statistical Association* 64(328):1183-1210. doi: 10.1080/01621459.1969.10501049.

Huynen, Philippe, Patrick Jeuniaux, Benjamin Mine, Eric Maes, et Luc Robert. 2024. *La base de données du Casier judiciaire central*. *Rapport de recherche*. Bruxelles: Institut National de Criminalistique et de Criminologie.

Huynen, Philippe, Benjamin Mine, Eric Maes, Patrick Jeuniaux, et Luc Robert. 2024. *Vers un moniteur belge de la récidive : jalons pour le développement d'un prototype basé sur le Casier judiciaire central*. *Rapport de recherche*. Bruxelles: Institut National de Criminalistique et de Criminologie.

Jeuniaux, Patrick, Benjamin Mine, et Isabelle Detry. 2022. *Le développement d'une base de données intégrée pour l'étude des trajectoires pénales des radicaux*. *Rapport de recherche*. 53. Bruxelles: Institut National de Criminalistique et de Criminologie.

Maes, Eric, Benjamin Mine, Patrick Jeuniaux, Shanty Sarief, Philippe Huynen, et Luc Robert. 2024. *SIDIS-Griffie databank*. Brussel: Nationaal Instituut voor Criminalistiek en Criminologie.

Mine, Benjamin, et Charlotte Vanneste. 2011. *Recherche relative aux conditions de faisabilité d'une articulation des bases de données statistiques sous la forme d'un datawarehouse*. *Rapport de recherche*. 25. Bruxelles: Institut national de criminalistique et de criminologie.

Newcombe, H. B., J. M. Kennedy, S. J. Axford, et A. P. James. 1959. « Automatic Linkage of Vital Records: Computers Can Be Used to Extract "Follow-up" Statistics of Families from Files of Routine Records ». *Science* 130(3381):954-59. doi: 10.1126/science.130.3381.954.

Robinson, Ian, Jim Webber, et Emil Eifrem. 2015. *Graph Databases: New Opportunities for Connected Data*. 2. ed. Beijing: O'Reilly.

Vanneste, Charlotte. 2012. « Vers une statistique criminelle "intégrée" : un si long chemin … ». P. 5-32 in *Les statistiques pénales belges à l'heure de l'informatisation. Enjeux et perspectives*, édité par C. Vanneste, F. Vesentini, J. Louette, et B. Mine. Academia Press.

A. Annexes

A.1. Quantité de liens établis

A.1.1. Phase 1 (de dégrossissage)

Tableau 24 – Nombre de liens créés dans la phase 1 de dégrossissage (selon l'orientation et la méthode)

Orientation des liens	M^{RRN}	M^{JUG}	M^{TRI}	M^{PHO}	M^{TRI-i}	M^{PHO-i}	TOTAL
Entre SIDIS et CJCS	/	17.524	728.873	919.451	25.476	26.671	1.717.995
Dans SIDIS	/	63.470	310.306	638.078	22.794	25.240	1.059.888
Dans CJCS	/	356.572	2.411.352	/	79.268	82.564	2.929.756
TOTAL	/	437.566	3.450.531	1.557.529	127.538	134.475	5.707.639

A.1.2. Phase 2 (d'affinage)

Tableau 25 – Nombre de liens dans la phase 2 d'affinage (selon l'orientation et la méthode)

Orientation des liens	M^{RRN}	M^{JUG}	M^{TRI}	M^{PHO}	M^{TRI-i}	M^{PHO-i}	TOTAL
Entre SIDIS et CJCS	53.134	17.158	490.086	548.234	20.613	21.247	1.150.472
Dans SIDIS	226	21.134	42.781	50.969	8.790	9.511	133.411
Dans CJCS	12.877	168.183	175.338	/	28.174	28.712	413.284
TOTAL	66.237	206.475	708.205	599.203	57.577	59.470	1.697.167

A.2. Temps d'exécution pour créer les liens d'intégration

L'exécution séquentielle des six méthodes, et pour les trois orientations des liens considérées, a pris environ deux jours et 20 heures dans l'environnement informatique utilisé (voir Tableau 18). Naturellement, ce temps d'exécution est en partie fonction de la puissance du système utilisé. Le fait de rapporter les durées d'exécution brutes ici permet toutefois de mieux se rendre compte des limites concrètes qui ont été rencontrées.

Cela donne aussi un point de référence pour améliorer le programme. En effet, dans le cadre de la présente analyse, certaines parties du programme n'ont pas été optimisées. Par exemple, les enregistrements qui ont déjà été comparés une fois ne sont pas mis en mémoire, de telle manière qu'à chaque fois qu'il faut les comparer à nouveau, l'information les concernant (i.e., les nœuds d'intégration) doit être extraite à nouveau de la base de données. Une version ultérieure du système pourrait améliorer ce point en plaçant l'ensemble des enregistrements à comparer en mémoire vive.

Enfin, ce rapport sur les durées d'exécution permet de comparer les durées selon différents points de vue, tels que l'orientation des liens (voir Tableau 26) et les méthodes utilisées pour tracer les liens (voir Tableau 27), ce qui donne une idée du coût relatifs des différentes activités et méthodes.

La durée d'exécution la plus longue est pour intra-CJCS (voir Tableau 26). Sa durée est toutefois sous-évaluée puisque, comme dit précédemment, la méthode M^{PHO} n'y a pas été appliquée. La durée la plus courte est pour intra-SIDIS, tandis que pour inter-SIDIS-CJCS elle est entre les deux. Cet ordre n'est pas surprenant dans la mesure où la durée d'exécution est corrélée au nombre d'enregistrements à comparer. Ce dernier est à priori le plus important pour intra-CJCS et le moins important pour intra-SIDIS, puisque le nombre d'enregistrements de personnes de SIDIS est inférieur à celui de CJCS. À posteriori, comme le révèle la durée d'exécution, c'est bien le cas.

Tableau 26 – Durées d'exécution par orientation des liens (toutes méthodes confondues) – de la plus lente à la plus rapide

intra-CJCS	inter-SIDIS-CJCS	intra-SIDIS	TOTAL
54h 31m	9h 33m	3h 52m	67h 56m

Pour comparer le temps d'exécution des méthodes, on se limite à un cas de figure pour lequel toutes les méthodes ont été utilisées, et notamment à celui qui nous intéresse le plus : inter-SIDIS-CJCS (voir Tableau 27).

Tableau 27 – Durées d'exécution par méthode (inter-SIDIS-CJCS) – de la plus lente à la plus rapide

M^{TRI}	M^{PHO}	M^{JUG}	M^{RRN}	M^{TRI-i}	M^{PHO-i}	TOTAL
3h 57m	3h 48m	1h 3m	16m	16m	13m	67h 56m

On voit que les méthodes utilisant les prénoms et noms dans le sens normal (M^{TRI}, M^{PHO}) sont les plus lentes : presque 4h. Lorsque les prénoms et noms sont inversés (M^{TRI-i}, M^{PHO-i}), elles tournent autour de 15 minutes, car il y a vraisemblablement beaucoup moins de cas de ce genre à investiguer. Le peu de cas qui existent sont trouvés relativement rapidement dans la base de données grâce l'exploration efficace du graphe. Pareillement, les quelques enregistrements qui ont le RRN en commun (un peu plus de 50.000) sont comparés les uns aux autres au bout de 16 minutes (méthode M^{RRN}).

A.3. Le poids des preuves

A.3.1. Liens établis au sein de SIDIS

Tableau 28 – Fréquence du poids des liens établis entre les nœuds de SIDIS selon la procédure utilisée pour établir ce lien (dates de jugement, RRN, trigrammes, etc.)

MAX(S)	M^{RRN}	M^{JUG}	M^{TRI}	M^{PHO}	M^{TRI-i}	M^{PHO-i}
40	84	119	91	97	1	2
39	39	3	6	5	0	0
38	53	6	8	8	0	0
37	46	1	0	0	0	0
36	3	0	0	0	0	0
35	1	0	0	0	0	0
34	0	2.775	2.592	2.639	74	85

33	0	229	251	255	19	20
32	0	972	1.117	1.254	275	295
31	0	547	872	946	286	297
30	0	15.115	35.726	41.840	7.465	8.086
29	0	17	85	88	10	8
28	0	1.084	1.664	3.149	509	571
27	0	1	1	1	0	0
26	0	68	7	7	2	2
25	0	0	5	5	0	0
24	0	58	19	18	0	0
23	0	3	13	13	1	1
22	0	21	105	284	50	53
21	0	3	4	4	0	0
20	0	53	212	353	98	91
19	0	1	2	2	0	0
18	0	3	0	0	0	0
17	0	0	0	0	0	0
16	0	9	1	1	0	0
15	0	0	0	0	0	0
14	0	1	0	0	0	0
13	0	0	0	0	0	0
12	0	36	0	0	0	0
11	0	0	0	0	0	0
10	0	9	0	0	0	0
1-9	0	?	?	?	?	?
TOTAL	226	21.134	42.781	50.969	8.790	9.511

A.3.2. Liens établis au sein de CJCS

Tableau 29 – Fréquence du poids des liens établis entre les nœuds de CJCS selon la procédure utilisée pour établir ce lien (dates de jugement, RRN, trigrammes, etc.)

MAX(S)	M^{RRN}	M^{JUG}	M^{TRI}	M^{PHO}	M^{TRI-i}	M^{PHO-i}
40	3.943	4.372	19.239	0	84	69
39	2.777	42	12.313	0	48	35
38	2.442	2.503	15.979	0	313	309
37	1.758	212	16.393	0	94	85
36	606	600	1.997	0	58	52
35	386	200	1.724	0	45	46
34	521	821	2.973	0	208	207
33	169	74	1.376	0	21	22
32	84	440	2.280	0	74	76
31	39	82	1.119	0	37	35
30	82	794	6.526	0	219	222
29	41	180	9.273	0	160	155
28	6	1.056	11.506	0	1.669	1.656
27	6	75	13.075	0	220	223

26	8	816	13.058	0	1.910	1.888
25	8	6	10.010	0	206	203
24	1	53	532	0	147	143
23	0	0	1.294	0	67	67
22	0	28	592	0	67	82
21	0	11	281	0	82	85
20	0	5.649	2.773	0	551	915
19	0	117	2.844	0	2.557	2.591
18	0	144.997	4.778	0	2.644	2.821
17	0	291	13.630	0	9.940	9.870
16	0	4.083	252	0	170	190
15	0	627	9.509	0	6.583	6.665
14	0	12	5	0	0	0
13	0	40	7	0	0	0
12	0	2	0	0	0	0
11	0	0	0	0	0	0
10	0	0	0	0	0	0
1-9	0	?	?	?	?	?
TOTAL	12.877	168.183	175.338	0	28.174	28.712

A.3.3. Liens établis entre SIDIS et CJCS

Tableau 30 – Fréquence du poids des preuves utilisées pour établir des liens entre SIDIS et CJCS, selon la procédure utilisée pour établir ce lien (dates de jugement, RRN, etc.)

MAX(S)	M^{RRN}	M^{JUG}	M^{TRI}	M^{PHO}	M^{TRI-i}	M^{PHO-i}
40	28.963	4.319	30.866	30.782	68	50
39	1.300	126	1.450	1.454	3	3
38	16.107	2.165	17.007	16.997	113	92
37	2.205	347	2.377	2.373	14	12
36	1.314	203	1.430	1.426	8	3
35	1.087	172	1.151	1.147	13	11
34	1.627	2.924	42.210	42.301	79	63
33	20	142	3.664	3.658	9	9
32	188	1.636	24.521	24.668	157	158
31	24	402	6.493	6.460	65	58
30	222	1.357	137.522	138.157	579	543
29	17	381	40.707	40.916	333	316
28	29	1.743	95.572	102.694	6.719	6.773
27	12	55	25.661	27.117	429	429
26	18	897	30.913	31.070	2.642	2.622
25	1	2	3.529	3.481	148	145
24	0	28	835	1.462	38	40
23	0	3	682	1.057	9	8
22	0	75	3.657	15.243	703	788
21	0	5	920	2.531	395	401
20	0	137	7.644	37.300	1.571	2.160

< 20	0	39	11.275	15.940	6.518	6.563
TOTAL	53.134	17.158	490.086	548.234	20.613	21.247

A.4. La similarité des liens

A.4.1. Liens établis au sein de SIDIS

Tableau 31 – Fréquence de la similarité (présentée en pourcentages arrondis au niveau de l'unité) des liens établis entre nœuds de SIDIS, selon la procédure utilisée pour établir ce lien (dates de jugement, RRN, trigrammes, etc.)

Similarité des liens	M^{RRN}	M^{JUG}	M^{TRI}	M^{PHO}	M^{TRI-i}	M^{PHO-i}
100%	0	23	3	3	0	0
99%	0	0	0	0	0	0
98%	0	0	2	2	0	0
97%	5	0	10	10	1	1
96%	0	2	29	29	25	25
95%	1	0	14	14	1	1
94%	0	1	7	7	0	0
93%	0	2	18	18	22	22
92%	3	1	5	5	2	2
91%	0	3	9	7	4	3
90%	4	3	70	69	35	34
89%	0	0	2	3	1	1
88%	0	13	27	28	4	5
87%	3	11	29	29	13	15
86%	2	45	28	29	2	5
85%	1	9	5	6	2	3
84%	0	4	14	10	5	5
83%	1	51	47	52	20	19
82%	5	9	20	14	3	2
81%	0	12	13	11	3	3
80%	3	51	113	90	30	21
79%	0	12	17	15	0	0
78%	0	24	15	18	6	5
77%	1	6	39	39	8	8
76%	0	52	159	219	45	39
75%	5	44	81	101	7	12
74%	1	7	30	30	1	1
73%	0	294	212	241	52	49
72%	3	7	54	57	33	33
71%	3	19	76	75	24	25
70%	4	349	1.060	1.932	168	206
69%	0	22	11	11	0	0
68%	0	28	35	57	8	5
67%	0	40	113	166	18	25
66%	0	383	1.623	1.633	681	676
65%	3	64	65	74	16	12
64%	0	198	188	207	56	68

63%	0	477	1.240	1.706	152	194
62%	0	61	81	86	11	13
61%	1	91	163	174	23	25
60%	1	4.002	4.667	4.670	605	599
59%	0	125	141	149	37	37
58%	1	279	244	256	35	36
57%	1	129	392	392	267	251
56%	1	5.409	2.144	2.016	259	337
55%	1	553	494	495	14	14
54%	0	387	284	372	95	107
53%	1	1.763	22.646	24.801	5.497	5.989
52%	4	1.482	135	147	3	10
51%	0	213	420	453	89	90
50%	1	4.374	5.487	9.941	407	478
49%	0	?	?	?	?	?
48%	4	?	?	?	?	?
47%	1	?	?	?	?	?
46%	1	?	?	?	?	?
45%	0	?	?	?	?	?
44%	2	?	?	?	?	?
43%	1	?	?	?	?	?
42%	0	?	?	?	?	?
41%	0	?	?	?	?	?
40%	1	?	?	?	?	?
39%	2	?	?	?	?	?
38%	0	?	?	?	?	?
37%	0	?	?	?	?	?
36%	3	?	?	?	?	?
35%	3	?	?	?	?	?
34%	0	?	?	?	?	?
33%	5	?	?	?	?	?
32%	0	?	?	?	?	?
31%	4	?	?	?	?	?
30%	3	?	?	?	?	?
29%	3	?	?	?	?	?
28%	11	?	?	?	?	?
27%	4	?	?	?	?	?
26%	3	?	?	?	?	?
25%	1	?	?	?	?	?
24%	15	?	?	?	?	?
23%	17	?	?	?	?	?
22%	57	?	?	?	?	?
21%	25	?	?	?	?	?
TOTAL	226	21.134	42.781	50.969	8.790	9.511

A.4.2. Liens établis au sein de CJCS

Tableau 32 – Fréquence de la similarité (présentée en pourcentages arrondis au niveau de l'unité) des liens établis entre nœuds de CJCS, selon la procédure utilisée pour établir ce lien (dates de jugement, RRN, trigrammes, etc.)

Similarité des liens	M^{RRN}	M^{JUG}	M^{TRI}	M^{PHO}	M^{TRI-i}	M^{PHO-i}
100%	8.593	5.706	10.173	0	228	231
99%	0	0	0	0	0	0
98%	663	563	807	0	3	3
97%	1.910	606	1.957	0	11	11
96%	64	167	1.180	0	308	308
95%	41	50	151	0	11	11
94%	224	165	333	0	29	32
93%	105	184	370	0	16	17
92%	956	954	1.448	0	48	44
91%	207	232	410	0	4	4
90%	23	211	480	0	25	25
89%	31	54	445	0	136	132
88%	9	155	752	0	279	279
87%	3	35	104	0	12	18
86%	14	59	315	0	48	47
85%	2	96	297	0	36	40
84%	7	69	407	0	113	107
83%	6	161	286	0	50	57
82%	5	46	1.768	0	1.504	1.508
81%	3	35	217	0	16	12
80%	2	95	2.056	0	1.154	1.138
79%	2	19	233	0	19	17
78%	1	49	223	0	28	27
77%	0	28	1.390	0	322	324
76%	0	118	1.463	0	216	137
75%	2	101	637	0	97	105
74%	1	8	110	0	21	13
73%	1	95	1.864	0	358	296
72%	0	112	649	0	78	64
71%	0	95	394	0	31	31
70%	0	346	4.113	0	1.696	1.688
69%	0	50	384	0	40	44
68%	0	45	1.043	0	97	87
67%	0	138	638	0	60	66
66%	0	5.153	5.371	0	1.975	1.983
65%	0	377	2.166	0	110	118
64%	0	77	3.325	0	553	561
63%	1	93	1.749	0	497	345
62%	0	522	2.333	0	73	74
61%	0	4.495	2.916	0	355	369
60%	0	4.285	3.771	0	661	745
59%	0	14	5.378	0	79	76
58%	0	1.610	9.416	0	5.957	5.911
57%	0	272	6.787	0	781	787

56%	0	711	8.882	0	161	155
55%	0	84.928	11.259	0	1.312	1.458
54%	0	20	2.500	0	70	72
53%	0	437	19.138	0	5.028	5.123
52%	0	1.458	11.843	0	2.172	2.404
51%	0	94	19.989	0	149	151
50%	0	52.790	21.418	0	1.147	1.457
(…)	0	?	?	?	?	?
23%	1	?	?	?	?	?
TOTAL	12.877	168.183	175.338	0	28.174	28.712

A.4.3. Liens établis entre SIDIS et CJCS

Tableau 33 – Fréquence de la similarité (présentée en pourcentages arrondis au niveau des dizaines) des liens établis entre SIDIS et CJCS, selon la procédure utilisée pour établir ce lien (dates de jugement, RRN, etc.)

SIM	M^{RRN}	M^{JUG}	M^{TRI}	M^{PHO}	M^{TRI-i}	M^{PHO-i}
100%	72	690	20.556	20.565	39	39
90%	41.154	10.444	195.438	195.488	767	772
80%	11.125	4.456	78.700	76.540	1.725	1.684
70%	615	586	19.044	17.687	2.755	2.724
60%	85	412	46.554	48.841	4.645	4.710
50%	30	570	129.794	189.113	10.682	11.318
40%	15	?	?	?	?	?
30%	21	?	?	?	?	?
20%	17	?	?	?	?	?
TOTAL	53.134	17.158	490.086	548.234	20.613	21.247

Tableau 34 – Fréquence de la similarité (présentée en pourcentages arrondis au niveau de l'unité) des liens établis entre SIDIS et CJCS, selon la procédure utilisée pour établir ce lien (dates de jugement, RRN, trigrammes, etc.)

Similarité des liens	M^{RRN}	M^{JUG}	M^{TRI}	M^{PHO}	M^{TRI-i}	M^{PHO-i}
100%	72	690	20.556	20.565	39	39
99%	0	0	0	0	0	0
98%	50	81	154	154	1	1
97%	5.005	884	7.414	7.433	10	10
96%	4.371	2.114	114.098	114.198	507	510
95%	4.895	1.018	11.334	11.353	27	28
94%	5.414	1.347	12.312	12.331	46	47
93%	1.765	1.065	14.402	14.395	76	76
92%	3.074	949	7.932	7.921	45	45
91%	1.598	452	4.166	4.163	9	9
90%	14.982	2.534	23.626	23.540	46	46
89%	7.322	1.237	19.555	19.449	214	212
88%	387	1.406	23.701	23.624	220	217
87%	522	690	10.154	10.151	33	32

86%	457	165	3.536	3.269	39	38
85%	689	221	3.375	3.151	94	88
84%	886	169	2.780	2.610	343	338
83%	130	101	3.628	3.030	57	54
82%	175	214	5.585	5.227	219	213
81%	156	121	2.252	2.199	119	115
80%	401	132	4.134	3.830	387	377
79%	33	43	768	654	20	17
78%	305	114	1.039	923	83	57
77%	40	35	2.104	2.030	473	463
76%	35	76	2.231	1.824	223	200
75%	68	147	2.911	2.731	128	129
74%	7	6	249	237	5	5
73%	73	93	2.035	1.892	325	333
72%	29	16	1.598	1.567	249	255
71%	15	18	855	837	104	100
70%	10	38	5.254	4.992	1.145	1.165
69%	11	21	521	505	58	54
68%	27	25	1.162	1.457	135	128
67%	6	21	2.064	2.267	183	195
66%	0	15	6.440	6.442	1.136	1.144
65%	7	28	2.156	3.102	179	187
64%	5	88	7.391	7.361	766	765
63%	16	21	3.807	3.795	729	747
62%	1	5	5.401	5.397	81	81
61%	1	27	2.181	2.979	434	425
60%	11	161	15.431	15.536	944	984
59%	2	21	7.503	8.469	154	163
58%	2	20	2.077	2.220	366	367
57%	11	45	21.578	23.092	3.832	3.859
56%	1	119	4.857	4.914	105	111
55%	2	105	16.270	17.287	808	870
54%	0	40	1.613	11.902	135	166
53%	0	87	45.134	46.337	1.698	1.709
52%	6	22	2.249	6.306	1.630	1.590
51%	2	5	7.038	8.957	143	139
50%	4	106	21.475	59.629	1.811	2.344
49%	0	?	?	?	?	?
48%	1	?	?	?	?	?
47%	6	?	?	?	?	?
46%	1	?	?	?	?	?
45%	0	?	?	?	?	?
44%	1	?	?	?	?	?
43%	3	?	?	?	?	?
42%	0	?	?	?	?	?
41%	2	?	?	?	?	?
40%	1	?	?	?	?	?
39%	1	?	?	?	?	?
38%	2	?	?	?	?	?
37%	4	?	?	?	?	?

36%	0	?	?	?	?	?
35%	1	?	?	?	?	?
34%	4	?	?	?	?	?
33%	0	?	?	?	?	?
32%	5	?	?	?	?	?
31%	1	?	?	?	?	?
30%	3	?	?	?	?	?
29%	1	?	?	?	?	?
28%	4	?	?	?	?	?
27%	1	?	?	?	?	?
26%	1	?	?	?	?	?
25%	1	?	?	?	?	?
24%	0	?	?	?	?	?
23%	7	?	?	?	?	?
22%	2	?	?	?	?	?
TOTAL	53.134	17.158	490.086	548.234	20.613	21.247

A.5. Seuils de poids et de similarité

A.5.1. Liens établis au sein de SIDIS

Tableau 35 – Nombre de liens établis au sein de SIDIS et figurant dans plusieurs fenêtres de poids et de similarité, selon les six méthodes.

Nombre de liens	**M^{RRN}**	**M^{JUG}**	**M^{TRI}**	**M^{PHO}**	**M^{TRI-i}**	**M^{PHO-i}**
Au total	226	21.134	42.781	50.969	8.790	9.511
Large [L]	45	1.005	2.207	3.162	517	545
En pourcentage	20%	5%	5%	6%	6%	6%
Moyenne [M]	28	189	437	407	162	156
En pourcentage	12%	1%	1%	1%	2%	2%
Étroite [E]	13	13	156	153	88	86
En pourcentage	6%	0%	0%	0%	1%	1%

Via le RRN

Tableau 36 – Distribution des liens établis entre les nœuds de SIDIS via le RRN en fonction de leur poids et similarité

similarité poids	100%	90%	80%	70%	60%	50%	40%	30%	20%	TOTAL
40	0	12	13	12	4	8	6	9	20	84
39	0	1	0	3	1	2	1	3	28	39
38	0	0	1	2	0	0	0	4	46	53

37	0	0	0	0	0	0	1	2	43	46
36	0	0	1	0	0	0	0	0	2	3
35	0	0	0	0	0	0	1	0	0	1
TOTAL	0	13	15	17	5	10	9	18	139	226

Via la date de jugement

Tableau 37 – Distribution des liens établis entre les nœuds de SIDIS via la date du jugement en fonction de leur poids et similarité

similarité / poids	100%	90%	80%	70%	60%	50%	TOTAL
40	0	1	1	11	17	89	119
39	0	0	0	1	1	1	3
38	0	0	0	1	1	4	6
37	0	0	0	0	0	1	1
34	0	4	32	43	111	2.585	2.775
33	0	1	11	3	23	191	229
32	0	2	11	19	91	849	972
31	0	0	7	29	95	416	547
30	1	4	112	639	4.664	9.695	15.115
29	0	0	0	2	5	10	17
28	0	0	1	48	263	772	1.084
27	0	0	0	0	0	1	1
26	0	0	1	3	49	15	68
24	0	0	1	10	24	23	58
23	0	0	0	0	2	1	3
22	0	0	2	2	6	11	21
21	0	0	0	0	1	2	3
20	0	0	2	0	4	47	53
19	0	0	0	1	0	0	1
18	0	0	0	0	2	1	3
16	2	0	2	1	4	0	9
14	1	0	0	0	0	0	1
12	11	0	21	1	3	0	36
10	8	0	1	0	0	0	9
TOTAL	23	12	205	814	5.366	14.714	21.134

Via les trigrammes

Tableau 38 – Distribution des liens établis entre les nœuds de SIDIS via les trigrammes en fonction de leur poids et similarité

similarité / poids	100%	90%	80%	70%	60%	50%	TOTAL
40	0	12	8	9	8	54	91
39	0	1	0	3	0	2	6
38	0	0	1	0	0	7	8
34	0	28	52	85	181	2.246	2.592

33	1	10	15	13	48	164	251
32	0	10	30	52	142	883	1.117
31	0	12	23	81	178	578	872
30	1	81	144	1.330	7.293	26.877	35.726
29	1	1	1	4	15	63	85
28	0	3	2	100	256	1.303	1.664
27	0	0	0	0	1	0	1
26	0	0	0	0	5	2	7
25	0	0	0	0	4	1	5
24	0	0	0	10	2	7	19
23	0	0	0	3	5	5	13
22	0	6	6	46	17	30	105
21	0	0	0	1	2	1	4
20	0	0	16	5	27	164	212
19	0	0	0	0	2	0	2
16	0	0	0	1	0	0	1
TOTAL	3	164	298	1.743	8.186	32.387	42.781

Via les représentations phonétiques

Tableau 39 – Distribution des liens établis entre les nœuds de SIDIS via les sons en fonction de leur poids et similarité

similarité / poids	100%	90%	80%	70%	60%	50%	TOTAL
40	0	12	10	11	9	55	97
39	0	1	0	3	0	1	5
38	0	0	0	0	0	8	8
34	0	26	48	83	189	2.293	2.639
33	1	10	13	12	52	167	255
32	0	10	25	60	188	971	1.254
31	0	12	26	86	207	615	946
30	1	80	123	2.287	7.756	31.593	41.840
29	1	1	2	4	14	66	88
28	0	3	2	115	310	2.719	3.149
27	0	0	0	0	1	0	1
26	0	0	0	0	5	2	7
25	0	0	0	0	4	1	5
24	0	0	0	10	2	6	18
23	0	0	0	3	5	5	13
22	0	6	6	46	17	209	284
21	0	0	0	1	2	1	4
20	0	0	17	5	21	310	353
19	0	0	0	0	2	0	2
16	0	0	0	1	0	0	1
TOTAL	3	161	272	2.727	8.784	39.022	50.969

Via les trigrammes inversés

Tableau 40 – Distribution des liens établis entre les nœuds de SIDIS via les trigrammes inversés en fonction de leur poids et similarité

similarité / poids	100%	90%	80%	70%	60%	50%	TOTAL
40	0	1	0	0	0	0	1
34	0	4	4	6	2	58	74
33	0	0	2	0	6	11	19
32	0	15	16	13	30	201	275
31	0	26	13	14	38	195	286
30	0	42	32	252	1.414	5.725	7.465
29	0	1	0	0	3	6	10
28	0	1	5	24	65	414	509
26	0	0	0	0	1	1	2
23	0	0	0	0	0	1	1
22	0	0	0	33	4	13	50
20	0	0	11	2	7	78	98
TOTAL	0	90	83	344	1.570	6.703	8.790

Via les représentations phonétiques inversées

Tableau 41 – Distribution des liens établis entre les nœuds de SIDIS via les sons inversés en fonction de leur poids et similarité

similarité / poids	100%	90%	80%	70%	60%	50%	TOTAL
40	0	1	0	0	1	0	2
34	0	3	5	6	5	66	85
33	0	0	1	1	5	13	20
32	0	15	16	15	32	217	295
31	0	25	14	14	42	202	297
30	0	42	27	281	1.440	6.296	8.086
29	0	1	0	0	1	6	8
28	0	1	5	26	85	454	571
26	0	0	0	0	1	1	2
23	0	0	0	0	0	1	1
22	0	0	0	33	2	18	53
20	0	0	11	2	3	75	91
TOTAL	0	88	79	378	1.617	7.349	9.511

A.5.2. Liens établis au sein de CJCS

Tableau 42 – Nombre de liens établis au sein de CJCS et figurant dans plusieurs fenêtres de poids et de similarité, selon les six méthodes.

Nombre de liens	**M^{RRN}**	**M^{JUG}**	**M^{TRI}**	**M^{PHO}**	**M^{TRI-i}**	**M^{PHO-i}**
Au total	12.877	168.183	175.338	/	28.174	28.712
Large [L]	12.875	10.196	25.998	/	1.611	1.550
En pourcentage	100%	6%	15%	/	6%	5%
Moyenne [M]	12.867	9.110	19.712	/	1.120	1.079
En pourcentage	100%	5%	11%	/	4%	4%
Étroite [E]	12.721	7.498	14.321	/	114	114
En pourcentage	99%	4%	8%	/	0%	0%

Via le RRN

Tableau 43 – Distribution des liens établis entre les nœuds de CJCS via le RRN en fonction de leur poids et similarité

similarité							
poids	100%	90%	80%	70%	60%	20%	TOTAL
40	2.242	1.696	5	0	0	0	3.943
39	2.286	486	4	1	0	0	2.777
38	1.266	1.132	41	2	1	0	2.442
37	1.452	295	10	1	0	0	1.758
36	380	222	4	0	0	0	606
35	214	168	4	0	0	0	386
34	420	94	4	2	0	1	521
33	141	24	3	1	0	0	169
32	54	28	2	0	0	0	84
31	32	7	0	0	0	0	39
30	53	29	0	0	0	0	82
29	28	8	5	0	0	0	41
28	4	2	0	0	0	0	6
27	5	1	0	0	0	0	6
26	8	0	0	0	0	0	8
25	7	1	0	0	0	0	8
24	1	0	0	0	0	0	1
TOTAL	8.593	4.193	82	7	1	1	12.877

Via la date de jugement

Tableau 44 – Distribution des liens établis entre les nœuds de CJCS via la date du jugement en fonction de leur poids et similarité

similarité / poids	100%	90%	80%	70%	60%	50%	TOTAL
40	2.242	1.017	0	192	315	606	4.372
39	16	24	0	2	0	0	42
38	1.256	700	37	63	153	294	2.503
37	87	44	1	8	27	45	212
36	328	180	4	12	34	42	600
35	68	85	4	4	20	19	200
34	463	206	40	40	21	51	821
33	10	31	9	12	5	7	74
32	107	211	39	27	12	44	440
31	8	40	16	10	0	8	82
30	164	211	138	85	62	134	794
29	41	58	34	14	19	14	180
28	410	168	132	94	119	133	1.056
27	6	15	16	11	11	16	75
26	249	83	106	94	133	151	816
25	1	3	2	0	0	0	6
24	17	4	15	11	4	2	53
22	3	10	5	2	5	3	28
21	0	3	1	0	1	6	11
20	69	29	12	237	3.883	1.419	5.649
19	0	0	1	1	9	106	117
18	145	10	177	51	9.489	135.125	144.997
17	2	0	1	0	4	284	291
16	9	0	11	0	398	3.665	4.083
15	2	0	2	0	510	113	627
14	1	0	1	0	0	10	12
13	2	0	0	0	1	37	40
12	0	0	1	1	0	0	2
TOTAL	5.706	3.132	805	971	15.235	142.334	168.183

Via les trigrammes

Tableau 45 – Distribution des liens établis entre les nœuds de CJCS via les trigrammes en fonction de leur poids et similarité

similarité / poids	100%	90%	80%	70%	60%	50%	TOTAL
40	2.241	1.694	10	266	1.642	13.386	19.239
39	2.284	485	5	91	619	8.829	12.313
38	1.256	1.115	45	99	693	12.771	15.979
37	1.440	288	10	74	461	14.120	16.393
36	376	219	8	26	79	1.289	1.997
35	213	165	13	36	141	1.156	1.724
34	445	262	121	351	251	1.543	2.973

33	225	257	67	106	151	570	1.376
32	102	275	95	67	158	1.583	2.280
31	113	294	66	66	150	430	1.119
30	159	413	383	840	1.192	3.539	6.526
29	145	172	272	471	2.042	6.171	9.273
28	408	527	632	497	1.726	7.716	11.506
27	119	159	226	649	1.556	10.366	13.075
26	249	452	673	638	2.986	8.060	13.058
25	159	124	251	354	2.442	6.680	10.010
24	15	23	34	75	74	311	532
23	10	64	32	62	315	811	1.294
22	1	15	45	142	159	230	592
21	0	8	17	85	98	73	281
20	14	37	436	540	573	1.173	2.773
19	8	1	192	446	883	1.314	2.844
18	10	78	109	1.209	2.554	818	4.778
17	98	1	1.598	3.225	715	7.993	13.630
16	0	4	3	119	73	53	252
15	82	3	1.304	539	1.958	5.623	9.509
14	1	0	0	2	0	2	5
13	0	1	0	1	5	0	7
TOTAL	10.173	7.136	6.647	11.076	23.696	116.610	175.338

Via les trigrammes inversés

Tableau 46 – Distribution des liens établis entre les nœuds de CJCS via les trigrammes inversés en fonction de leur poids et similarité

similarité poids	100%	90%	80%	70%	60%	50%	TOTAL
40	1	0	9	20	6	48	84
39	1	1	2	15	1	28	48
38	9	18	9	18	21	238	313
37	2	5	6	26	7	48	94
36	1	3	1	17	10	26	58
35	4	0	1	12	10	18	45
34	5	5	5	30	25	138	208
33	1	4	2	1	6	7	21
32	0	6	6	9	7	46	74
31	0	7	3	3	9	15	37
30	2	39	40	14	26	98	219
29	1	16	32	11	31	69	160
28	25	96	217	52	195	1.084	1.669
27	8	14	18	12	34	134	220
26	13	193	251	65	378	1.010	1.910
25	16	14	8	7	54	107	206
24	1	7	6	13	14	106	147
23	1	3	2	6	12	43	67
22	0	0	1	9	17	40	67
21	0	0	1	13	16	52	82
20	1	3	52	60	169	266	551

19	2	0	92	185	535	1.743	2.557
18	1	18	35	332	1.005	1.253	2.644
17	86	0	1.498	1.780	411	6.165	9.940
16	0	3	0	45	30	92	170
15	47	0	1.051	111	1.392	3.982	6.583
TOTAL	228	455	3.348	2.866	4.421	16.856	28.174

Via les représentations phonétiques inversées

Tableau 47 – Distribution des liens établis entre les nœuds de CJCS via les sons inversés en fonction de leur poids et similarité

similarité / poids	100%	90%	80%	70%	60%	50%	TOTAL
40	1	0	5	13	6	44	69
39	1	1	1	7	1	24	35
38	9	18	7	18	22	235	309
37	2	5	3	23	6	46	85
36	1	3	1	14	8	25	52
35	4	0	1	12	9	20	46
34	5	5	6	26	24	141	207
33	1	4	2	2	6	7	22
32	0	6	4	10	6	50	76
31	0	7	3	4	8	13	35
30	2	39	39	15	26	101	222
29	1	16	32	6	33	67	155
28	25	95	212	52	207	1.065	1.656
27	8	14	17	12	35	137	223
26	13	192	231	64	387	1.001	1.888
25	16	14	7	7	52	107	203
24	1	7	5	15	14	101	143
23	1	2	1	7	12	44	67
22	0	0	1	9	17	55	82
21	0	0	1	13	15	56	85
20	1	3	54	65	215	577	915
19	2	0	93	183	378	1.935	2.591
18	1	20	42	335	1.023	1.400	2.821
17	86	0	1.503	1.700	419	6.162	9.870
16	0	4	14	45	31	96	190
15	50	0	1.053	45	1.432	4.085	6.665
TOTAL	231	455	3.338	2.702	4.392	17.594	28.712

A.5.3. Liens établis entre SIDIS et CJCS

Tableau 48 – Nombre de liens établis entre SIDIS et CJCS et figurant dans plusieurs fenêtres de poids et de similarité, selon les six méthodes.

Nombre de liens	M^{RRN}	M^{JUG}	M^{TRI}	M^{PHO}	M^{TRI-i}	M^{PHO-i}
Original	53.134	17.158	490.086	548.234	20.613	21.247
Large [L]	52.966	16.150	307.897	304.421	3.016	2.923
En pourcentage	100%	94%	63%	56%	15%	14%
Moyenne [M]	52.351	15.490	287.818	285.718	1.545	1.507
En pourcentage	99%	90%	59%	52%	7%	7%
Étroite [E]	41.178	9.188	136.381	136.382	162	163
En pourcentage	77%	54%	28%	25%	1%	1%

Via le RRN

Tableau 49 – Distribution des liens établis entre SIDIS et CJCS via le RRN en fonction de leur poids et similarité

similarité / poids	100%	90%	80%	70%	60%	50%	40%	30%	20%	TOTAL
40	18	27.391	1.431	88	4	13	6	10	2	28.963
39	19	657	599	20	3	0	1	1	0	1.300
38	11	7.880	7.746	383	58	12	2	6	9	16.107
37	7	1.618	529	41	2	3	3	1	1	2.205
36	6	1.154	132	15	4	0	1	1	1	1.314
35	1	884	183	13	3	0	1	0	2	1.087
34	3	1.330	249	31	9	2	1	1	1	1.627
33	3	6	9	1	1	0	0	0	0	20
32	1	55	120	11	0	0	0	0	1	188
31	0	11	10	3	0	0	0	0	0	24
30	0	123	93	4	1	0	0	1	0	222
29	2	8	3	4	0	0	0	0	0	17
28	0	15	14	0	0	0	0	0	0	29
27	0	10	2	0	0	0	0	0	0	12
26	1	12	4	1	0	0	0	0	0	18
25	0	0	1	0	0	0	0	0	0	1
TOTAL	72	41.154	11.125	615	85	30	15	21	17	53.134

Via la date de jugement

Tableau 50 – Distribution des liens établis entre SIDIS et CJCS via la date du jugement en fonction de leur poids et similarité

similarité / poids	100%	90%	80%	70%	60%	50%	TOTAL
40	18	3.929	230	105	9	28	4.319
39	2	68	49	4	0	3	126
38	10	1.062	946	121	11	15	2.165
37	2	280	42	18	1	4	347
36	5	174	16	4	2	2	203
35	0	134	30	7	1	0	172
34	16	1.363	1.421	52	29	43	2.924
33	0	68	67	1	2	4	142
32	6	835	712	51	20	12	1.636
31	4	298	74	8	2	16	402
30	109	805	125	52	99	167	1.357
29	25	225	89	25	8	9	381
28	244	759	446	72	164	58	1.743
27	0	42	10	2	0	1	55
26	232	347	169	44	55	50	897
25	0	1	1	0	0	0	2
24	3	11	11	1	0	2	28
23	0	3	0	0	0	0	3
22	5	10	5	5	4	46	75
21	1	3	0	1	0	0	5
20	3	17	8	7	2	100	137
18	5	10	5	6	3	10	39
TOTAL	690	10.444	4.456	586	412	570	17.158

Via les trigrammes

Tableau 51 – Distribution des liens établis entre SIDIS et CJCS via les 3 premières lettres des noms et prénoms en fonction de leur poids et similarité

similarité / poids	100%	90%	80%	70%	60%	50%	TOTAL
40	18	27.348	1.803	855	68	774	30.866
39	19	657	599	55	5	115	1.450
38	10	7.852	7.754	832	117	442	17.007
37	7	1.616	525	109	9	111	2.377
36	6	1.149	156	58	11	50	1.430
35	1	880	179	57	7	27	1.151
34	16	15.490	18.493	693	1.057	6.461	42.210
33	22	1.032	1.009	57	230	1.314	3.664
32	7	9.418	10.733	871	639	2.853	24.521
31	23	3.377	1.472	207	339	1.075	6.493
30	112	67.321	4.362	3.343	17.366	45.018	137.522
29	12.329	4.261	2.097	1.101	5.931	14.988	40.707
28	247	36.958	14.855	2.447	9.802	31.263	95.572

27	6.920	1.973	1.006	1.019	2.998	11.745	25.661
26	232	12.823	9.057	1.290	2.377	5.134	30.913
25	358	836	400	172	496	1.267	3.529
24	2	219	202	203	115	94	835
23	28	326	62	73	116	77	682
22	2	293	742	1.043	544	1.033	3.657
21	55	71	118	292	127	257	920
20	3	870	1.604	705	1.027	3.435	7.644
19	117	5	514	461	822	1.314	3.233
18	5	659	521	1.344	1.921	455	4.905
17	16	0	363	1.746	116	472	2.713
16	0	2	6	9	6	0	23
15	0	2	68	2	308	20	400
14	1	0	0	0	0	0	1
TOTAL	20.556	195.438	78.700	19.044	46.554	129.794	490.086

Via les représentations phonétiques

Tableau 52 – Distribution des liens établis entre SIDIS et CJCS via les sons des noms et prénoms en fonction de leur poids et similarité

similarité							
poids	100%	90%	80%	70%	60%	50%	TOTAL
40	18	27.367	1.704	834	64	795	30.782
39	19	656	600	55	6	118	1.454
38	11	7.869	7.737	804	114	462	16.997
37	7	1.618	522	104	9	113	2.373
36	6	1.150	151	57	13	49	1.426
35	1	882	176	56	6	26	1.147
34	16	15.513	18.451	566	1.010	6.745	42.301
33	22	1.031	998	47	227	1.333	3.658
32	7	9.438	10.720	797	663	3.043	24.668
31	23	3.375	1.437	184	340	1.101	6.460
30	112	67.241	3.601	2.807	17.287	47.109	138.157
29	12.332	4.246	1.882	969	5.912	15.575	40.916
28	248	36.993	14.367	2.240	10.112	38.734	102.694
27	6.924	1.967	916	962	2.975	13.373	27.117
26	232	12.857	8.730	1.129	2.452	5.670	31.070
25	358	833	354	157	471	1.308	3.481
24	2	219	201	200	127	713	1.462
23	28	326	49	55	111	488	1.057
22	2	293	745	1.050	851	12.302	15.243
21	55	72	119	316	185	1.784	2.531
20	3	871	1.602	727	1.953	32.144	37.300
19	117	5	514	463	857	4.711	6.667
18	5	662	524	1.346	2.656	780	5.973
17	16	0	366	1.751	127	616	2.876
16	0	2	6	9	6	0	23
15	0	2	68	2	307	21	400
14	1	0	0	0	0	0	1
TOTAL	20.565	195.488	76.540	17.687	48.841	189.113	548.234

Via les trigrammes inversés

Tableau 53 – Distribution des liens établis entre SIDIS et CJCS via les 3 premières lettres des noms et prénoms inversés en fonction de leur poids et similarité

similarité / poids	100%	90%	80%	70%	60%	50%	TOTAL
40	0	4	10	19	34	1	68
39	0	0	1	1	1	0	3
38	0	14	30	17	46	6	113
37	0	0	9	0	4	1	14
36	0	0	0	2	6	0	8
35	0	3	1	1	4	4	13
34	0	12	18	7	30	12	79
33	0	3	2	0	2	2	9
32	0	25	50	19	30	33	157
31	1	20	11	4	14	15	65
30	1	79	64	44	201	190	579
29	2	43	32	31	94	131	333
28	5	252	318	200	1.075	4.869	6.719
27	13	14	14	31	81	276	429
26	6	229	239	81	498	1.589	2.642
25	5	11	4	13	35	80	148
24	0	2	2	2	1	31	38
23	0	2	1	0	2	4	9
22	0	5	98	213	238	149	703
21	2	3	55	158	34	143	395
20	0	14	310	134	626	487	1.571
19	3	0	313	298	503	1.556	2.673
18	0	32	26	465	818	708	2.049
17	1	0	116	1.015	28	395	1.555
16	0	0	0	0	1	0	1
15	0	0	1	0	239	0	240
TOTAL	39	767	1.725	2.755	4.645	10.682	20.613

Via les représentations phonétiques inversées

Tableau 54 – Distribution des liens établis entre SIDIS et CJCS via les sons des noms et prénoms inversés en fonction de leur poids et similarité

similarité / poids	100%	90%	80%	70%	60%	50%	TOTAL
40	0	4	2	22	21	1	50
39	0	0	1	0	2	0	3
38	0	14	26	15	32	5	92
37	0	0	7	0	4	1	12
36	0	0	0	1	2	0	3
35	0	3	1	1	2	4	11
34	0	13	14	5	20	11	63

33	0	3	1	0	2	3	9
32	0	25	49	15	22	47	158
31	1	20	10	3	9	15	58
30	1	79	55	34	179	195	543
29	2	43	27	27	83	134	316
28	5	253	321	168	1.085	4.941	6.773
27	13	14	13	28	85	276	429
26	6	231	232	69	506	1.578	2.622
25	5	11	2	17	31	79	145
24	0	2	2	1	1	34	40
23	0	2	0	0	2	4	8
22	0	5	98	213	289	183	788
21	2	3	55	158	43	140	401
20	0	14	310	146	698	992	2.160
19	3	0	312	301	503	1.509	2.628
18	0	33	28	467	821	766	2.115
17	1	0	117	1.033	28	400	1.579
16	0	0	0	0	1	0	1
15	0	0	1	0	239	0	240
TOTAL	39	772	1.684	2.724	4.710	11.318	21.247

A.6. Valeurs ajoutées des méthodes de liaison

A.6.1. Fenêtre de paramètres étroite

Tableau 55 – Distribution des liens selon les six méthodes de liaison et les bases de données concernées (fenêtre de paramètres étroite)

ID	M^{JUG}	M^{RRN}	M^{TRI}	M^{TRI-i}	M^{PHO}	M^{PHO-i}	inter SIDIS-CJCS	intra SIDIS	Intra CJCS	TOTAL
1	X	X	X	X	X	X	2	0	/	2
2	X	X	X	X	X		0	0	/	0
3	X	X	X	X		X	0	0	17	17
4	X	X	X	X			0	0	3	3
5	X	X	X		X	X	0	0	/	0
6	X	X	X		X		5.859	0	/	5.859
7	X	X	X			X	0	0	7	7
8	X	X	X				2	0	6.350	6.352
9	X	X		X	X	X	0	0	/	0
10	X	X		X	X		0	0	/	0
11	X	X		X		X	0	0	0	0
12	X	X		X			0	0	0	0
13	X	X			X	X	0	0	/	0
14	X	X			X		13	0	/	13
15	X	X				X	0	0	0	0
16	X	X					6	1	153	160
17	X		X	X	X	X	5	0	/	5

18	X		X	X	X		0	0	/	0
19	X		X	X		X	0	0	3	3
20	X		X	X			0	0	2	2
21	X		X		X	X	0	0	/	0
22	X		X		X		3.290	6	/	3.296
23	X		X			X	0	0	3	3
24	X		X				5	1	855	861
25	X			X	X	X	0	0	/	0
26	X			X	X		0	0	/	0
27	X			X		X	0	0	0	0
28	X			X			0	0	0	0
29	X				X	X	1	0	/	1
30	X				X		4	0	/	4
31	X					X	0	0	0	0
32	X						1	5	105	111
33		X	X	X	X	X	17	0	/	17
34		X	X	X	X		0	0	/	0
35		X	X	X		X	0	0	6	6
36		X	X	X			0	0	3	3
37		X	X		X	X	0	0	/	0
38		X	X		X		35.130	0	/	35.130
39		X	X			X	0	0	2	2
40		X	X				23	0	5.943	5.966
41		X		X	X	X	0	0	/	0
42		X		X	X		0	0	/	0
43		X		X		X	0	0	0	0
44		X		X			0	0	0	0
45		X			X	X	0	0	/	0
46		X			X		60	0	/	60
47		X				X	0	0	1	1
48		X					66	12	244	322
49			X	X	X	X	53	3	/	56
50			X	X	X		0	0	/	0
51			X	X		X	0	0	3	3
52			X	X			0	0	0	0
53			X		X	X	0	0	/	0
54			X		X		91.862	142	/	92.004
55			X			X	0	0	0	0
56			X				133	4	1.124	1.261
57				X	X	X	0	0	/	0
58				X	X		0	0	/	0
59				X		X	85	59	51	195
60				X			0	26	26	52
61					X	X	0	0	/	0
62					X		86	2	/	88
63						X	0	24	21	45

						TOTAL	136.703	285	14.922	151.910

A.6.2. Fenêtre de paramètres moyenne

Tableau 56 – Distribution des liens selon les six méthodes de liaison et les bases de données concernées (fenêtre de paramètres moyenne)

ID	M^{JUG}	M^{RRN}	M^{TRI}	M^{TRI-i}	M^{PHO}	M^{PHO-i}	inter SIDIS-CJCS	intra SIDIS	Intra CJCS	TOTAL
1	X	X	X	X	X	X	3	0	/	3
2	X	X	X	X	X		4	0	/	4
3	X	X	X	X		X	0	0	21	21
4	X	X	X	X			0	0	8	8
5	X	X	X		X	X	0	0	/	0
6	X	X	X		X		7.106	0	/	7.106
7	X	X	X			X	0	0	12	12
8	X	X	X				23	0	6.374	6.397
9	X	X		X	X	X	0	0	/	0
10	X	X		X	X		0	0	/	0
11	X	X		X		X	0	0	0	0
12	X	X		X			0	0	0	0
13	X	X			X	X	0	0	/	0
14	X	X			X		17	0	/	17
15	X	X				X	0	0	0	0
16	X	X					29	2	164	195
17	X		X	X	X	X	42	0	/	42
18	X		X	X	X		3	0	/	3
19	X		X	X		X	0	0	13	13
20	X		X	X			0	0	16	16
21	X		X		X	X	0	0	/	0
22	X		X		X		8.110	17	/	8.127
23	X		X			X	0	0	10	10
24	X		X				56	5	2.196	2.257
25	X			X	X	X	0	0	/	0
26	X			X	X		0	0	/	0
27	X			X		X	1	0	0	1
28	X			X			0	0	0	0
29	X				X	X	1	0	/	1
30	X				X		30	1	/	31
31	X					X	0	0	0	0
32	X						65	164	296	525
33		X	X	X	X	X	46	0	/	46
34		X	X	X	X		14	0	/	14
35		X	X	X		X	0	0	6	6
36		X	X	X			0	0	7	7

37		X	X		X	X	3	0	/	3
38		X	X		X		44.567	0	/	44.567
39		X	X			X	0	0	2	2
40		X	X				176	0	6.012	6.188
41		X		X	X	X	0	0	/	0
42		X		X	X		0	0	/	0
43		X		X		X	2	0	0	2
44		X		X			0	0	0	0
45		X			X	X	0	0	/	0
46		X			X		107	0	/	107
47		X				X	0	0	1	1
48		X					254	26	268	548
49			X	X	X	X	372	3	/	375
50			X	X	X		19	2	/	21
51			X	X		X	2	0	26	28
52			X	X			0	3	11	14
53			X		X	X	5	0	/	5
54			X		X		224.200	329	/	224.529
55			X			X	1	0	8	9
56			X				3.066	78	4.990	8.134
57				X	X	X	2	0	/	2
58				X	X		0	0	/	0
59				X		X	991	96	618	1.705
60				X			44	58	394	496
61					X	X	4	3	/	7
62					X		1.063	52	/	1.115
63						X	32	54	362	448
						TOTAL	290.460	893	21.815	313.168

A.6.3. Fenêtre de paramètres large

Tableau 57 – Distribution des liens selon les six méthodes de liaison et les bases de données concernées (fenêtre de paramètres large)

ID	M^{JUG}	M^{RRN}	M^{TRI}	M^{TRI-i}	M^{PHO}	M^{PHO-i}	inter SIDIS-CJCS	intra SIDIS	Intra CJCS	TOTAL
1	X	X	X	X	X	X	6	0	/	6
2	X	X	X	X	X		5	0	/	5
3	X	X	X	X		X	0	0	24	24
4	X	X	X	X			0	0	18	18
5	X	X	X		X	X	2	0	/	2
6	X	X	X		X		7.139	0	/	7.139
7	X	X	X			X	0	0	17	17
8	X	X	X				30	0	6.359	6.389
9	X	X		X	X	X	0	0	/	0

10	X	X		X	X		0	0	/	0
11	X	X		X		X	0	0	0	0
12	X	X		X			0	0	0	0
13	X	X			X	X	0	0	/	0
14	X	X			X		17	0	/	17
15	X	X				X	0	0	0	0
16	X	X					45	3	164	212
17	X		X	X	X	X	46	0	/	46
18	X		X	X	X		6	0	/	6
19	X		X	X		X	0	0	18	18
20	X		X	X			0	0	18	18
21	X		X		X	X	0	0	/	0
22	X		X		X		8.532	46	/	8.578
23	X		X			X	0	0	11	11
24	X		X				86	8	2.610	2.704
25	X			X	X	X	0	0	/	0
26	X			X	X		0	0	/	0
27	X			X		X	1	1	0	2
28	X			X			0	0	0	0
29	X				X	X	1	0	/	1
30	X				X		47	17	/	64
31	X					X	0	0	0	0
32	X						187	930	957	2.074
33		X	X	X	X	X	57	0	/	57
34		X	X	X	X		38	0	/	38
35		X	X	X		X	0	0	11	11
36		X	X	X			0	0	25	25
37		X	X		X	X	25	0	/	25
38		X	X		X		44.840	0	/	44.840
39		X	X			X	0	0	10	10
40		X	X				239	0	5.985	6.224
41		X		X	X	X	0	0	/	0
42		X		X	X		0	0	/	0
43		X		X		X	4	0	0	4
44		X		X			0	1	0	1
45		X			X	X	0	0	/	0
46		X			X		115	0	/	115
47		X				X	0	0	2	2
48		X					404	41	268	713
49			X	X	X	X	549	16	/	565
50			X	X	X		78	2	/	80
51			X	X		X	6	0	32	38
52			X	X			9	3	16	28
53			X		X	X	12	0	/	12
54			X		X		240.265	1.252	/	241.517
55			X			X	1	0	14	15

56			X				5.926	880	10.830	17.636
57				X	X	X	2	0	/	2
58				X	X		0	0	/	0
59				X		X	2.096	328	891	3.315
60				X			113	166	558	837
61					X	X	8	5	/	13
62					X		2.631	1.824	/	4.455
63						X	107	196	521	824
						TOTAL	313.675	5.719	29.359	348.753

A.7. Statistiques relatives aux nœuds de personnes

Tableau 58 – Nombre de personnes correspondant à des cas ambigus (24 scénarios)

	Type de fenêtre		**[L]**	**[M]**	**[ME/RRN]**	**[ME]**	**[E]**	**[RRN]**
	Type de dossiers CJCS	**Liens inter ou intra**						
1	actif & inactif	inter & intra	18.768	9.328	7.589	7.482	3.822	3.162
2	actif & inactif	inter	8.075	2.576	1.975	1.868	270	110
3	actif	inter & intra	8.590	2.894	5.156	2.241	456	111
4	actif	inter	8.024	2.538	1.946	1.840	263	110

A.7.1. Scénarios 1-6

Tableau 59 – Nombre de personnes selon la source de ses enregistrements et selon qu'elles ont été condamnées définitivement à une peine d'emprisonnement ou pas (scénarios 1-6)

	CJCS	**SIDIS**	**Peine prison**	**[L]**	**[M]**	**[ME/RRN]**	**[ME]**	**[E]**	**[RRN]**
1	X			3.530.275	3.553.263	3.578.200	3.578.877	3.708.872	3.793.103
2	X	X	X	190.075	183.762	169.147	168.710	88.934	37.971
3	X	X		105.584	101.750	93.597	93.393	46.811	14.746
4		X		46.430	52.702	60.867	61.122	108.271	140.489
5		X	X	17.913	25.802	40.438	40.932	121.070	172.089
			TOTAL	3.890.277	3.917.279	3.942.249	3.943.034	4.073.958	4.158.398

Tableau 60 – Statistiques additionnelles relatives à SIDIS et CJCS (scénarios 1-6)

	Statistique	**[L]**	**[M]**	**[ME/RRN]**	**[ME]**	**[E]**	**[RRN]**
A	Nombre de personnes dans SIDIS	360.002	364.016	364.049	364.157	365.086	365.295
B	Nombre de personnes de SIDIS qui sont aussi dans CJCS	295.659	285.512	262.744	262.103	135.745	52.717
C	Proportion de personnes de SIDIS qui sont aussi dans CJCS (C = B / A)	82,13%	78,43%	72,17%	71,98%	37,18%	14,43%
D	Nombre de personnes ayant été condamnées définitivement à une peine d'emprisonnement	207.988	209.564	209.585	209.642	210.004	210.060

E	Proportion de personnes de SIDIS ayant été condamnées définitivement à une peine d'emprisonnement (E = D / A)	57,77%	57,57%	57,57%	57,57%	57,52%	57,50%
F	Nombre de personnes ayant été condamnées définitivement à une peine d'emprisonnement et étant dans CJCS	190.075	183.762	169.147	168.710	88.934	37.971
G	Proportion de personnes ayant été condamnées définitivement à une peine d'emprisonnement et étant dans CJCS (G = F/ D)	91,39%	87,69%	80,71%	80,48%	42,35%	18,08%

A.7.2. Scénarios 7-12 : Uniquement inter

Tableau 61 – Nombre de personnes selon la source de ses enregistrements et selon qu'elles ont été condamnées définitivement à une peine d'emprisonnement ou pas (scénarios 7-12 : inter)

	CJCS	SIDIS	Peine prison	[L]	[M]	[ME/RRN]	[ME]	[E]	[RRN]
1	X			3.543.693	3.561.649	3.585.099	3.585.780	3.713.569	3.796.833
2	X	X	X	190.522	183.805	169.189	168.748	88.932	37.971
3	X	X		105.888	101.800	93.639	93.435	46.813	14.746
4		X		47.291	52.958	61.153	61.406	108.414	140.494
5		X	X	18.220	25.859	40.527	41.015	121.125	172.100
			TOTAL	3.905.614	3.926.071	3.949.607	3.950.384	4.078.853	4.162.144

Tableau 62 – Statistiques additionnelles relatives à SIDIS et CJCS (scénarios 7-12 : inter)

	Statistique	[L]	[M]	[ME/RRN]	[ME]	[E]	[RRN]
A	Nombre de personnes dans SIDIS	361.921	364.422	364.508	364.604	365.284	365.311
B	Nombre de personnes de SIDIS qui sont aussi dans CJCS	296.410	285.605	262.828	262.183	135.745	52.717
C	Proportion de personnes de SIDIS qui sont aussi dans CJCS (C = B / A)	81,90%	78,37%	72,10%	71,91%	37,16%	14,43%
D	Nombre de personnes ayant été condamnées définitivement à une peine d'emprisonnement	208.742	209.664	209.716	209.763	210.057	210.071
E	Proportion de personnes de SIDIS ayant été condamnées définitivement à une peine d'emprisonnement (E = D / A)	57,68%	57,53%	57,53%	57,53%	57,51%	57,50%
F	Nombre de personnes ayant été condamnées définitivement à une peine d'emprisonnement et étant dans CJCS	190.522	183.805	169.189	168.748	88.932	37.971
G	Proportion de personnes ayant été condamnées définitivement à une peine d'emprisonnement et étant dans CJCS (G = F/ D)	91,27%	87,67%	80,68%	80,45%	42,34%	18,08%

A.7.3. Scénarios 13-18 : Uniquement actifs

Tableau 63 – Nombre de personnes selon la source de ses enregistrements et selon qu'elles ont été condamnées définitivement à une peine d'emprisonnement ou pas (scénarios 13-18 : actifs)

	CJCS	SIDIS	Peine prison	[L]	[M]	[ME/RRN]	[ME]	[E]	[RRN]

1	X			3.468.490	3.486.226	3.506.298	3.509.791	3.637.078	3.720.046
2	X	X	X	189.503	183.160	168.871	168.427	88.801	37.966
3	X	X		105.312	101.458	93.473	93.266	46.743	14.745
4		X		46.770	53.010	60.998	61.255	108.341	140.490
5		X	X	18.576	26.431	40.732	41.233	121.204	172.094
			TOTAL	3.828.651	3.850.285	3.870.372	3.873.972	4.002.167	4.085.341

Tableau 64 – Statistiques additionnelles relatives à SIDIS et CJCS (scénarios 13-18 : actifs)

	Statistique	**[L]**	**[M]**	**[ME/RRN]**	**[ME]**	**[E]**	**[RRN]**
A	Nombre de personnes dans SIDIS	360.161	364.059	364.074	364.181	365.089	365.295
B	Nombre de personnes de SIDIS qui sont aussi dans CJCS	294.815	284.618	262.344	261.693	135.544	52.711
C	Proportion de personnes de SIDIS qui sont aussi dans CJCS (C = B / A)	81,86%	78,18%	72,06%	71,86%	37,13%	14,43%
D	Nombre de personnes ayant été condamnées définitivement à une peine d'emprisonnement	208.079	209.591	209.603	209.660	210.005	210.060
E	Proportion de personnes de SIDIS ayant été condamnées définitivement à une peine d'emprisonnement (E = D / A)	57,77%	57,57%	57,57%	57,57%	57,52%	57,50%
F	Nombre de personnes ayant été condamnées définitivement à une peine d'emprisonnement et étant dans CJCS	189.503	183.160	168.871	168.427	88.801	37.966
G	Proportion de personnes ayant été condamnées définitivement à une peine d'emprisonnement et étant dans CJCS (G = F/ D)	91,07%	87,39%	80,57%	80,33%	42,29%	18,07%

A.7.4. Scénarios 19-24 : Uniquement inter et actifs

Tableau 65 – Nombre de personnes selon la source de ses enregistrements et selon qu'elles ont été condamnées définitivement à une peine d'emprisonnement ou pas (scénarios 19-24 : inter / actifs)

	CJCS	**SIDIS**	**Peine prison**	**[L]**	**[M]**	**[ME/RRN]**	**[ME]**	**[E]**	**[RRN]**
1	X			3.468.490	3.486.226	3.509.111	3.509.791	3.637.078	3.720.046
2	X	X	X	189.888	183.199	168.913	168.465	88.799	37.966
3	X	X		105.569	101.503	93.515	93.308	46.745	14.745
4		X		47.638	53.268	61.286	61.541	108.484	140.495
5		X	X	18.892	26.492	40.824	41.319	121.261	172.105
			TOTAL	3.830.477	3.850.688	3.873.649	3.874.424	4.002.367	4.085.357

Tableau 66 – Statistiques additionnelles relatives à SIDIS et CJCS (scénarios 19-24 : inter / actifs)

	Statistique	**[L]**	**[M]**	**[ME/RRN]**	**[ME]**	**[E]**	**[RRN]**
A	Nombre de personnes dans SIDIS	361.987	364.462	364.538	364.633	365.289	365.311
B	Nombre de personnes de SIDIS qui sont aussi dans CJCS	295.457	284.702	262.428	261.773	135.544	52.711
C	Proportion de personnes de SIDIS qui sont aussi dans CJCS (C = B / A)	81,62%	78,12%	71,99%	71,79%	37,11%	14,43%

D	Nombre de personnes ayant été condamnées définitivement à une peine d'emprisonnement	208.780	209.691	209.737	209.784	210.060	210.071
E	Proportion de personnes de SIDIS ayant été condamnées définitivement à une peine d'emprisonnement (E = D / A)	57,68%	57,53%	57,54%	57,53%	57,51%	57,50%
F	Nombre de personnes ayant été condamnées définitivement à une peine d'emprisonnement et étant dans CJCS	189.888	183.199	168.913	168.465	88.799	37.966
G	Proportion de personnes ayant été condamnées définitivement à une peine d'emprisonnement et étant dans CJCS (G = F/ D)	90,95%	87,37%	80,54%	80,30%	42,27%	18,07%

Collection des rapports et notes de recherche
Collectie van onderzoeksrapporten en onderzoeksnota's

Actualisée en juin 2024 – Geactualiseerd in juni 2024

N°59 MAES, E., MINE, B., JEUNIAUX, P., SARIEF, S., HUYNEN, P., ROBERT, L., (2024), SIDIS-Griffie databank, Onderzoeksrapport, Nationaal Instituut voor Criminalistiek en Criminologie, Operationele Directie Criminologie, Collectie van onderzoeksrapporten en onderzoeksnota's, 103 p.

N°58 HUYNEN, P., JEUNIAUX, P., MINE, B., MAES, E., ROBERT, L., (2024), La base de données du casier judiciaire central, Rapport de recherche, Institut National de Criminalistique et de Criminologie. Direction Opérationnelle de Criminologie. Collection des rapports et notes de recherche, 127 p.

N°57b BURSSENS, D., (2023), Tendances de la criminalité. Le crime drop au niveau international et en Belgique, Rapport de recherche, Institut National de Criminalistique et de Criminologie. Direction Opérationnelle de Criminologie. Collection des rapports et notes de recherche, 67 p.

N°57a BURSSENS, D., (2023), Trends in Criminaliteit. De crime drop internationaal en in België, Onderzoeksrapport, Nationaal Instituut voor Criminalistiek en Criminologie, Operationele Directie Criminologie, Collectie van onderzoeksrapporten en onderzoeksnota's, 63 p.

N°56b BAUWENS, A., SCHILS, E., LEMONNE, A. (prom.), RAVIER, I. (prom.), (2023), Verkennend onderzoek in verband met de invoering van een methodologie voor de retrospectieve analyse van feminicides in België, Onderzoeksrapport, Nationaal Instituut voor Criminalistiek en Criminologie, Operationele Directie Criminologie, Collectie van onderzoeksrapporten en onderzoeksnota's, 64 p.

N°56a BAUWENS, A., SCHILS, E., LEMONNE, A. (prom.), RAVIER, I. (prom.), (2023), Recherche exploratoire portant sur la mise en place d'une méthodologie d'analyse rétrospective des cas de féminicide en Belgique, Rapport de recherche, Institut National de Criminalistique et de Criminologie. Direction Opérationnelle de Criminologie. Collection des rapports et notes de recherche, 64 p.

N°55b REMACLE, C., DETRY, I., MINE, B., JEUNIAUX, P., (2023), De sociaaljuridische trajecten van terugkeerders in België : stand van zaken van de betrokkene actoren en van de bestaande procedures. Onderzoeksrapport, Nationaal Instituut voor Criminalistiek en Criminologie, Operationele Directie Criminologie, Collectie van onderzoeksrapporten en onderzoeksnota's, 84 p.

N°55 REMACLE, C., DETRY, I., MINE, B., JEUNIAUX, P., (2023), Les parcours socio-judiciaires des returnees en Belgique : état des lieux des acteurs impliqués et des procédures mises en place. Rapport de recherche de l'Institut National de Criminalistique et de Criminologie, Direction Opérationnelle Criminologie, Collection des rapports et notes de recherche, 86 p.

N°54 BRUYERE, L., TANGE, C., (2021), Recherche exploratoire portant sur les représentations des policiers et pratiques policières en matière de reportabilité des faits relevant du « harcèlement de rue ». Rapport de recherche de l'Institut National de Criminalistique et de Criminologie, Direction Opérationnelle Criminologie, Collection des rapports et notes de recherche, 31 p.

N°53 JEUNIAUX, P., MINE B, DETRY, I. (2022), Le développement d'une base de données intégrée pour l'étude des trajectoires pénales des radicaux. Rapport de recherche de l'Institut National de Criminalistique et de Criminologie, Direction Opérationnelle Criminologie, Collection des rapports et notes de recherche, 234 p.

N°52 VARGA, R., VANNESTE C. (dir) (2022), L'incidence de la politique antiterroriste belge sur l'application du droit des étrangers. A travers la jurisprudence du Conseil du contentieux des étrangers (CCE). Rapport de la recherche réalisée dans le cadre du programme AFFECT (Evaluation de l'impact des politiques belges de déradicalisation sur la cohésion sociale et les libertés) financé par BELSPO (volet CCE), Collection des rapports de recherche de la Direction opérationnelle de Criminologie n°52, Institut National de Criminalistique et de Criminologie, 110 p. (décembre 2022)

N°51a REMACLE C., VANNESTE C. (dir), VAN PRAET S. (2022) Approche ethnographique et jurisprudentielle des poursuites en matière de terrorisme en Belgique. Rapport de la recherche réalisée dans le cadre du programme AFFECT (Evaluation de l'impact des politiques belges de déradicalisation sur la cohésion sociale et les libertés) financé par BELSPO (volet judiciaire) », *Collection des rapports de recherche de la Direction opérationnelle de Criminologie* n°51, Institut National de Criminalistique et de Criminologie, 340 p. + Rapport 51b (synthèse)

N°50 MINE, B., JEUNIAUX, P., DETRY, I.. (2022) La radicalité verbalisée. Analyse du discours de personnes radicales à propos de leur engagement et de leur(s) expérience(s) avec les autorités. Rapport de la recherche. Projet financé par la Politique scientifique fédérale (BELSPO), *Collection des rapports de recherche de la Direction opérationnelle de Criminologie* n°50, Institut National de Criminalistique et de Criminologie, 210 p.

N°49 JONCKHEERE, A., SCHILS, E., *La médiation SAC en temps de COVID sur le territoire des 19 communes de la Région de Bruxelles-Capitale. Etude réalisée en 2021-2022 dans le cadre de la recherche « Les sanctions administratives communales dans le cadre des mesures anti-COVID : administration de la justice pénale et respect des droits fondamentaux »*, Nationaal Instituut voor Criminalistiek en Criminologie/Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie/Operationele Directie Criminologie, Bruxelles/Brussel, octobre 2022, 62 p.

N°48c RAVIER, I., VAN PRAET, S., *Les dossiers judiciaires : la gestion du costume pénal de l'IPV. Analyse des dossiers., BELSPO,* Nationaal Instituut voor Criminalistiek en Criminologie/Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie/Operationele Directie Criminologie, *Belspo*, Bruxelles/Brussel, mai 2022, 122 p.

N°48a VANNESTE, C., *Violences entre partenaires : Impact, processus, évolution et politiques publiques. Analyse des entretiens menés avec des acteurs-clé du secteur policier et de l'assistance policière aux victimes en Fédération Wallonie-Bruxelles.* Nationaal Instituut voor Criminalistiek en Criminologie/Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie/Operationele Directie Criminologie, *Belspo*, IPV Pro&Pol, Bruxelles/Brussel, décembre 2022, 148 p.

N°47 DETRY, I., MINE, B., JEUNIAUX, P., *La radicalisation au prisme des banques de données. Rapport de recherche dans le cadre du projet FAR. Projet financé par BELSPO,* Nationaal Instituut voor Criminalistiek en Criminologie/Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie/Operationele Directie Criminologie, KU Leuven, ULB, Bruxelles/Brussel, avril 2021, 65 p.

N°46 MAHIEU, V., TANGE, C.(PROM), SMEETS, S, (PROM.) *Projet de recherche portant sur le partage de l'espace public à Schaerbeek (PEPS). Projet financé par la zone de police Shaerbeek-Evere-St-Josse (POLBRUNO),* Nationaal Instituut voor Criminalistiek en Criminologie/Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie/Operationele Directie Criminologie, Centre de recherches Pénalité, sécurité & déviance, Bruxelles/Brussel, septembre 2019, 25 p.

N°45 GOTELAERE, S., SCHILS, E., JONCKHEERE, A, (PROM.) *Recherche portant sur les pratiques en matière de médiation dans le cadre des sanctions administratives communales,* Nationaal Instituut voor Criminalistiek en Criminologie/Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie/Operationele Directie

Criminologie, SPP Intégration Sociale / POD Maatschappelijke Integratie, Bruxelles/Brussel, novembre/november 2020, 117 p.

N°44b MAHIEU, V., VAN PRAET, DETRY, I., (PROM.), TANGE C., (PROM.) *Een analyse van geseponeerde dossiers met een tenlastelegging inzake de discriminatiewetgeving,* Nationaal Instituut voor Criminalistiek en Criminologie/Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie/Operationele Directie Criminologie, Unia, Fondation Roi Baudouin, Bruxelles/Brussel, janvier/januari 2021, 51 p.

N°44a MAHIEU, V., VAN PRAET, DETRY, I., (PROM.), TANGE C., (PROM.) *Une analyse des dossiers judiciaires classes sans suite comprenant une prévention liée à la discrimination,* Nationaal Instituut voor Criminalistiek en Criminologie/Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie/Operationele Directie Criminologie, Unia, Fondation Roi Baudouin, Bruxelles/Brussel, novembre/november 2020, 50 p.

N°43c VAN PRAET, S., TANGE, C. (PROM.), *Identifying and tackling problematic or abusive forms of police selectivity. An action research on the problematic practices and/or mechanisms of police selectivity in the police district of Schaerbeek-Evere-St-Josse (PolBruNo),* Nationaal Instituut voor Criminalistiek en Criminologie/Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie/Operationele Directie Criminologie, Unia, PolBruno, Bruxelles/Brussel, juillet/juli 2020, 74 p.

N°43b VAN PRAET, S., TANGE, C. (PROM.), *Identificeren en aanpakken van problemen of misbruiken bij politieselectiviteit. Een actiononderzoek naar problematische praktijken en mechanismes van politieselectiviteit in de politiezone Schaarbeek-Evere-Sint-Joost-ten-Node (PolBruNo),* Nationaal Instituut voor Criminalistiek en Criminologie/Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie/Operationele Directie Criminologie, Unia, PolBruno, Bruxelles/Brussel, juillet/juli 2020, 80 p.

N°43a VAN PRAET, S., TANGE, C. (PROM.), *Identifier et affronter des problèmes et abus dans la sélectivité policière. Une recherche-action sur les pratiques et/ou mécanismes problématiques de sélectivité policière au sein de la zone de police schaerbeek-Evere-St-Josse (PolBruNo),* Nationaal Instituut voor Criminalistiek en Criminologie/Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie/Operationele Directie Criminologie, Unia, PolBruno, Bruxelles/Brussel, juillet/juli 2020, 79 p.

N°42 DE BLANDER, R., ROBERT, L., MINCKE, C., MAES, E., MINE, B., *Etude de faisabilité d'un moniteur de la récidive / Haalbaarheidsstudie betreffende een recidivemonitor,* Nationaal Instituut voor Criminalistiek en Criminologie/Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie/Operationele Directie Criminologie, Bruxelles/Brussel, Mai/Mei 2019, 44 p.

N°41 VANNESTE, C., *La politique criminelle en matière de violences conjugales : une évaluation des pratiques judiciaires et de leurs effets en termes de récidive,* Nationaal Instituut voor Criminalistiek en Criminologie/Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie/Operationele Directie Criminologie, Bruxelles/Brussel, Mai/Mei 2016, 131 p.

VANNESTE, C., *Het strafrechtelijk beleid op het vlak van partnergeweld : een evaluatie van de rechtspraktijk en de gevolgen ervan inzake recidive,* Nationaal Instituut voor Criminalistiek en Criminologie/Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie/Operationele Directie Criminologie, Bruxelles/Brussel, Mai/Mei 2016, 135 p.

N°40 MAHIEU, V., RAVIER, I., VANNESTE, C., *Vers une image chiffrée de la délinquance enregistrée des jeunes en Région de Bruxelles-Capitale / Naar een beeldvorming van geregistreerde delinquentie bij jongeren in het Brussels Hoofdstedelijk Gewest,* Nationaal Instituut voor Criminalistiek en Criminologie/Institut National de Criminalistique et de

Criminologie, Direction Opérationnelle de Criminologie/Operationele Directie Criminologie, Bruxelles/Brussel, Juin 2015, 154 p.

N°39 BURSSENS, D., TANGE, C., MAES, E., *Op zoek naar determinanten van de toepassing en de duur van de voorlopige hechtenis. A la recherche de déterminants du recours à la détention préventive et de sa durée.,* Nationaal Instituut voor Criminalistiek en Criminologie/Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie/Operationele Directie Criminologie, Bruxelles/Brussel, Juni/juin 2015, 103 p.

N°38 MINE, B., ROBERT, L., *Recidive na een rechterlijke beslissing. Nationale cijfers op basis van het Centraal Strafregister. La récidive après une décision judiciaire. Des chiffres nationaux sur la base du Casier judiciaire central..,* Nationaal Instituut voor Criminalistiek en Criminologie/Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie/Operationele Directie Criminologie, Bruxelles/Brussel, Mai 2015, 62 p.

N°37 RAVIER, I., *l'évolution des signalements de mineurs pour faits qualifiés infraction : quelles pistes de compréhension ?.,* Nationaal Instituut voor Criminalistiek en Criminologie/Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie/Operationele Directie Criminologie, Bruxelles/Brussel, Mai 2015, 56 p.

N°36 JONCKHEERE, A., *Le rôle et l'organisation des greffiers d'instruction.,* Nationaal Instituut voor Criminalistiek en Criminologie/Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie/Operationele Directie Criminologie, Bruxelles/Brussel, Septembre 2014, 76 p.

N°35 MAHIEU, V., LEMONNE, A. (dir.), VANNESTE, C. (dir.), *Projet de recherche portant sur le développement d'un outil d'aide à la décision en matière de violences entre partenaires. Projet réalisé dans le cadre d'une collaboration avec l'équipe de l'Institut Thomas More Kempen.*, Nationaal Instituut voor Criminalistiek en Criminologie/Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie/Operationele Directie Criminologie, Bruxelles/Brussel, Avril 2014, 99 p.

N°34 DACHY, A., BOLIVAR, D., LEMONNE, A. (dir.), VANNESTE, C. (dir.), *Implementing a better response to victims' needs. Handbook accomplished in the framework of the project « Restorative justice, Urban Security and Social Inclusion : a new European approach » JUST/2010/JPEN/1601. Financed by CRIMINAL JUSTICE Programme EU 2008-2010*, Nationaal Instituut voor Criminalistiek en Criminologie/Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie/Operationele Directie Criminologie, Bruxelles/Brussel, 2012, 103 p.

N°33 MINE, B., ROBERT, L., JONCKHEERE, A. (DIR.), MAES, E. (dir.), *Analyse des processus de travail de la Direction Gestion de la détention et des directions pénitentiaires locales dans le cadre de la formulation d'avis et de la prise de décisions en matière de modalités d'exécution des peines/Analyse van werkprocessen van de Directie Detentiebeheer en lokale gevangenisdirecties in het kader van de advies- en besluitvorming inzake bijzondere strafuitvoeringsmodaliteiten*, Nationaal Instituut voor Criminalistiek en Criminologie/Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie/Operationele Directie Criminologie, Bruxelles/Brussel, février/februari 2013, 370 p.

N°32b GILBERT, E., MAHIEU, V., GOEDSEELS, E. (PROM.), RAVIER, I. (PROM.), *Onderzoek naar de beslissingen van jeugdrechters/jeugdrechtbanken in MOF-zaken*, Nationaal Instituut voor Criminalistiek en Criminologie, Operationele Directie Criminologie, Onderzoeksrapport, Brussel, september 2012, 189 p.

N°32a GILBERT, E., MAHIEU, V., GOEDSEELS, E. (DIR.), RAVIER, I. (DIR.), *Recherche relative aux décisions des juges/tribunaux de la jeunesse dans les affaires de faits qualifiés infractions,*

Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie, Rapport final de recherche, Bruxelles, septembre 2012, 189 p.

N°31 MAHIEU, V., VANDERSTRAETEN, B., LEMONNE, A. (dir.), *Evaluation du Forum national pour une politique en faveur des victimes/ Evaluatie van het Nationaal Forum voor Slachtofferbeleid*. *Rapport final/Eindrapport(bilingue)*, Nationaal Instituut voor Criminalistiek en Criminologie/Institut National de Criminalistique et de Criminologie, Operationele Directie Criminologie/Direction Opérationnelle de Criminologie, Brussel/Bruxelles, février/februari 2012, 220 p + annexes.

N°30 ADELAIRE K., REYNAERT J.-F., NISEN L., *Recherche relative au système de rémunération de l'aide juridique de deuxième ligne*, MINCKE C., SHOENAERS F. (dir.), Centre de recherche et d'interventions sociologiques de l'Université de Liège / Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie, Bruxelles, septembre 2012, 156 p + annexes.

N°29 JEUNIAUX, P, RENARD, B. (dir), *Les dépenses en matière d'expertises génétiques dans le système pénal belge, de 2000 à 2010*, Institut National de Criminalistique et de Criminologie, Rapport final de recherche, Bruxelles, janvier 2012, 185 p.

N°28 JONCKHEERE, A., *La (mise en) liberté sous conditions : usages et durée d'une mesure alternative à la détention préventive (2005-2009). Note de recherche dans le cadre de l'exploitation scientifique de SIPAR, la base de données des maisons de justice*, Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie, Bruxelles, février 2012, 12p.

N°27 ROBERT, L., MAES, E. (dir.), *Wederopsluiting na vrijlating uit de gevangenis*, Nationaal Instituut voor Criminalistiek en Criminologie, Operationele Directie Criminologie, Brussel, 27 januari 2012, 151p. + bijl.

N°26 DEVRESSE (dir.), M.-S., ROBERT, L., VANNESTE, C. (dir.), coll. HELLEMANS, A., *Onderzoek inzake de classificatie van en de vraag naar regimes binnen de strafinrichtingen/Recherche relative à la classification et à la question des régimes au sein des établissements pénitentiaires*, Nationaal Instituut voor Criminalistiek en Criminologie/Institut National de Criminalistique et de Criminologie, Operationele Directie Criminologie/Direction Opérationnelle de Criminologie, Brussel/Bruxelles, 2011, 276 p.

N°25 MINE, B., VANNESTE, C. (dir.), *Recherche relative aux conditions de faisabilité d'une articulation des bases de données statistiques sous la forme d'un « Datawarehouse »*, Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie, Bruxelles, décembre 2011, 220 p.

N°24b BURSSENS, D., VANNESTE, C. (dir.), *La médiation pénale. Note de recherche dans le cadre de l'exploitation scientifique de SIPAR, la base de données des maisons de justice*, Institut National de Criminalistique et de Criminologie, Direction Opérationnelle de Criminologie, Bruxelles, mai 2011, 38 p.

N°24a BURSSENS, D., VANNESTE, C. (dir.), *Bemiddeling in strafzaken. Onderzoeksnota in het kader van de wetenschappelijke exploitatie van SIPAR, databank van de justitiehuizen*, Nationaal Instituut voor Criminalistiek en Criminologie, Operationele Directie Criminologie, Brussel, mei 2011, 38 p.

N°23 DE MAN, C., MAES, E. (dir.), MINE, B., VAN BRAKEL, R., *Toepassingsmogelijkheden van het elektronisch toezicht in het kader van de voorlopige hechtenis – Possibilités d'application de la surveillance électronique dans le cadre de la détention préventive*, Eindrapport - Rapport final, Brussel/Bruxelles, Nationaal Instituut voor Criminalistiek en Criminologie/Institut National de Criminalistique et de Criminologie, Département de Criminologie, Operationele Directie Criminologie/Direction Opérationnelle de Criminologie, december/décembre 2009, 304 p. + bijlagen/annexes.

N° 22 HEYLEN B., RAVIER I., SCHOFFELEN J., VANNESTE C. (dir.), *Une recherche évaluative d'un centre fermé pour mineurs, le centre « De Grubbe » à Everberg/Evaluatieonderzoek van een gesloten instelling voor jongeren, centrum « De Grubbe » te Everberg*, Rapport final/Eindrapport, Institut National de Criminalistique et de Criminologie, Département de Criminologie/Nationaal Instituut voor Criminalistiek en Criminologie, Hoofdafdeling Criminologie, Bruxelles/Brussel, 2009, 193 p.

N° 21b JONCKHEERE A., VANNESTE C. (dir.), *Wetenschappelijke exploitatie van SIPAR, de databank van de justitiehuizen. Analyse van de gegevens betreffende het jaar 2006*, Nationaal Instituut voor Criminalistiek en Criminologie, Hoofdafdeling Criminologie, Brussel, februari 2009, 111 p.

N° 21 JONCKHEERE A., VANNESTE C. (dir.), *Recherche relative à l'exploitation scientifique de SIPAR, la base de données des maisons de justice. Analyse de données relatives à l'année 2006*, Institut National de Criminalistique et de Criminologie, Département de Criminologie, Bruxelles, juillet 2008, 141 p.

N° 20b GOEDSEELS E., DETRY I., VANNESTE C. (dir.), *Recherche relative à l'exploitation scientifique des données disponibles en matière de protection de la jeunesse et de délinquance juvénile, Premier rapport, Analyse du flux des affaires entrées au niveau des parquets de la jeunesse en 2005*, Institut National de Criminalistique et de Criminologie, Département de Criminologie, Bruxelles, juillet 2007, 112 p. + annexes.

N° 20a GOODSEELS E., DETRY I., VANNESTE C. (dir.), *Onderzoek met betrekking tot de productie en wetenschappelijke exploitatie van cijfergegevens aangaande jeugddelinquentie en jeugdbescherming, Eerste onderzoeksrapport, Analyse van de instroom op de jeugdparketten voor het jaar 2005*, Nationaal Instituut voor Criminalistiek en Criminologie, Hoofdafdeling Criminologie, Brussel, juli 2007, 116 p. + bijlagen.

N° 19b LEMONNE A., VAN CAMP T., VANFRAECHEM I., VANNESTE C. (dir.), *Onderzoek met betrekking tot de evaluatie van de voorzieningen ten behoeve van slachtoffers van inbreuken*, Eindrapport, Nationaal Instituut voor Criminalistiek en Criminologie, Hoofdafdeling Criminologie, Brussel, juli 2007, 356 p. + bijlagen.

N° 19a LEMONNE A., VAN CAMP T., VANFRAECHEM I., VANNESTE C. (dir.), *Recherche relative à l'évaluation des dispositifs mis en place à l'égard des victimes d'infraction*, Rapport final, Institut National de Criminalistique et de Criminologie, Département de Criminologie, Bruxelles, juillet 2007, 354 p. + annexes.

N° 18 MAES E., i.s.m. het Directoraat-generaal Uitvoering van Straffen en Maatregelen (DELTENRE, S. en VAN DEN BERGH, W.), *Strafbecijfering en -uitvoering in België anno 2006. Analyse van de actuele praktijk en voorstelling van enkele alternatieve denkpistes*, Onderzoeksnota, Nationaal Instituut voor Criminalistiek en Criminologie, Hoofdafdeling Criminologie, Brussel, 26 september 2006, 37 p. + bijlagen.

N° 17 MAES E., *Proeve van werklastmeting van de toekomstige strafuitvoeringsrechtbanken. Een simulatie-oefening op basis van data in verband met de strafuitvoeringspraxis tijdens het jaar 2004*, Onderzoeksnota, Nationaal Instituut voor Criminalistiek en Criminologie, Hoofdafdeling Criminologie, Brussel, 13 december 2005 (met aanvulling d.d. 19 mei 2006: tabel in bijlage), 10 p. + bijlagen.

N° 16b JONCKHEERE A., VANNESTE C. (dir.), *Onderzoek met betrekking tot de wetenschappelijke exploitatie van het gegevensbestand betreffende de justitiehuizen – SIPAR*, Eerste rapport (vertaling uit het Frans), Nationaal Instituut voor Criminalistiek en Criminologie, Hoofdafdeling Criminologie, Brussel, december 2006, 83 p.

N° 16a JONCKHEERE A., VANNESTE C. (dir.), *Recherche relative à l'exploitation scientifique des bases de données existantes au sein des Maisons de justice – SIPAR*, Premier rapport, Institut National de Criminalistique et de Criminologie, Département de Criminologie, Bruxelles, décembre 2006, 77 p.

N° 15b RENARD B., VANNESTE C. (dir.), *Het statuut van de deskundige in strafzaken*, Eindrapport, Nationaal Instituut voor Criminalistiek en Criminologie, Hoofdafdeling Criminologie, Brussel, december 2005, (gedeeltelijke vertaling, april 2006), 86 p.

N° 15a RENARD B., VANNESTE C. (dir.), *Le statut de l'expert en matière pénale*, Rapport final de recherche, Institut National de Criminalistique et de Criminologie, Département de Criminologie, Bruxelles, décembre 2005, 405 p.

N° 14 GOOSSENS F., MAES E., DELTENRE S., VANNESTE C. (dir.), *Projet de recherche relatif à l'introduction de la surveillance électronique comme peine autonome/Onderzoeksproject inzake de invoering van het elektronisch toezicht als autonome straf*, Rapport final de recherche/Eindrapport, Institut National de Criminalistique et de Criminologie, Département de Criminologie/Nationaal Instituut voor Criminalistiek en Criminologie, Hoofdafdeling Criminologie, Bruxelles/Brussel, octobre/oktober 2005, 204 p. + bijlagen/annexes.

N° 13 DAENINCK P., DELTENRE S., JONCKHEERE A., MAES E., VANNESTE C. (dir.), *Analyse des moyens juridiques susceptibles de réduire la détention préventive/Analyse van de juridische mogelijkheden om de toepassing van de voorlopige hechtenis te verminderen*, Rapport final de recherche/Eindrapport, Institut National de Criminalistique et de Criminologie, Département de Criminologie/Nationaal Instituut voor Criminalistiek en Criminologie, Hoofdafdeling Criminologie, Bruxelles/Brussel, mars/maart 2005, 367 p.

N° 12 RENARD B., DELTENRE S., *L'expertise en matière pénale – Phase 1: Cartographie des pratiques*, Institut National de Criminalistique et de Criminologie, Département de Criminologie, Rapport final de recherche, Bruxelles, juin 2003, 138 p. + annexes.

N° 11 DELTENRE S., MAES E., *Analyse statistique sur base de données de condamnations: plus-value et applications concrètes/Statistische analyse aan de hand van de veroordelingsgegevens: meerwaarde en praktijkvoorbeeld*, Notes de recherche/Onderzoeksnota's, Institut National de Criminalistique et de Criminologie, Département de Criminologie/Nationaal Instituut voor Criminalistiek en Criminologie, Hoofdafdeling Criminologie, Bruxelles/Brussel, 2000-2002.

N° 10 MAES E., *Studie van de evolutie van de gedetineerdenpopulatie volgens misdrijfcategorie (1980-1998)*, Onderzoeksnota, Nationaal Instituut voor Criminalistiek en Criminologie, Hoofdafdeling Criminologie, Brussel, september 2001, 15 p. + bijlagen.

N° 9 DELTENRE S., MAES E., *Effectmeting van enkele mogelijke wetswijzigingen op het vlak van de voorlopige hechtenis/Simulations de l'impact de quelques modifications législatives en matière de détention préventive*, Onderzoeksnota's/Notes de recherche, Nationaal Instituut voor Criminalistiek en Criminologie, Hoofdafdeling Criminologie/Institut National de Criminalistique et de Criminologie, Département de Criminologie, Brussel/Bruxelles, 2001.

N° 8b VANNESTE C., *De beslissingen genomen door de parketmagistraten en de jeugdrechters ten aanzien van delinquente minderjarigen*, Eindrapport (vertaling), Nationaal Instituut voor Criminalistiek en Criminologie, Hoofdafdeling Criminologie, Brussel, dec. 2001, 206 p. + bijlagen.

N° 8a VANNESTE C., *Les décisions prises par les magistrats du parquet et les juges de la jeunesse à l'égard des mineurs délinquants*, Rapport final de recherche, Institut National de Criminalistique et de Criminologie, Département de Criminologie, Bruxelles, juin 2001, 205 p. + annexes.

N° 7 RENARD B., *L'usage du polygraphe en procédure pénale; analyse procédurale, Note d'étude – Partie III de l'avis pour le Ministre de la Justice et le Collège des Procureurs généraux sur l'usage du polygraphe en procédure pénale belge*, Institut National de Criminalistique et de Criminologie, Département de Criminologie, Bruxelles, septembre 2000, 59-80 p.

N° 6 MAES E., DUPIRE V., TORO F., VANNESTE C. (dir.), *De V.I.-commissies in actie. Onderzoek naar de werking van de in het kader van de nieuwe V.I.-wetgeving (wetten van 5 en 18 maart*

1998) opgerichte commissies voor de voorwaardelijke invrijheidstelling/Les commissions de libération conditionnelle en action. Recherche sur le fonctionnement des commissions de libération conditionnelle créées dans le cadre de la nouvelle réglementation sur la libération conditionnelle (lois des 5 et 18 mars 1998), Eindrapport/Rapport final de recherche, Nationaal Instituut voor Criminalistiek en Criminologie, Hoofdafdeling Criminologie/Institut National de Criminalistique et de Criminologie, Département de Criminologie, Brussel/Bruxelles, augustus/août 2000, 355 p. + bijlagen/annexes.

N° 5 MORMONT, C. (DIR.), VANNESTE, C. (DIR.), TORO, F., MARSDEN, E., SNIJDERS, J., *Etude comparative dans les 15 pays de l'Union Européenne relative au statut et modalités de l'expertise des personnes présumées ou avérées abuseurs sexuels*, Rapport final de la recherche co-financée par la Commission Européenne et le Ministère de la Justice belge, Programme européen STOP, Université de Liège et Institut National de Criminalistique et de Criminologie, Département de Criminologie, octobre 1999, 192 p. + résumés en néerlandais (11 p.) et anglais (11 p).

N° 4 RENARD B., VANDERBORGHT J., *Recherche Proactive, révélateur d'une approche nouvelle? Etude relative à la recherche proactive dans le cadre de la lutte contre la criminalité organisée Proactieve Recherche, exponent van een vernieuwde aanpak? Onderzoek naar de proactieve recherche in de strijd tegen de georganiseerde criminaliteit*, Institut National de Criminalistique et de Criminologie, Département de Criminologie/Nationaal Instituut voor Criminalistiek en Criminologie, Hoofdafdeling Criminologie, Rapport final de recherche/Eindrapport, Bruxelles/Brussel, septembre/september 1999, 386 p.

N° 3 SNACKEN S. (dir.), DELTENRE S., RAES A., VANNESTE C., VERHAEGHE P., *Recherche qualitative sur l'application de la détention préventive et de la liberté sous conditions/Kwalitatief onderzoek naar de toepassing van de voorlopige hechtenis en de vrijheid onder voorwaarden*, Rapport final de recherche/Eindrapport, Institut National de Criminalistique et de Criminologie, Département de Criminologie/Nationaal Instituut voor Criminalistiek en Criminologie, Hoofdafdeling Criminologie/Vrije Universiteit Brussel, Bruxelles/Brussel, 1999, 244 p.

N° 2 SNACKEN S. (dir.), DE BUCK K., D'HAENENS K., RAES A., VERHAEGHE P., *Onderzoek naar de toepassing van de voorlopige hechtenis en de vrijheid onder voorwaarden*, Eindrapport, Nationaal Instituut voor Criminalistiek en Criminologie, Hoofdafdeling Criminologie/Vrije Universiteit Brussel, Brussel, 1997, 174 p.

N° 1 DE BUCK K., D'HAENENS K., *Electronic Monitoring*, Studienota, Nationaal Instituut voor Criminalistiek en Criminologie, Hoofdafdeling Criminologie, 1996, 40 p.

Direction Opérationnelle de Criminologie
Operationele Directie Criminologie

TOUR DES FINANCES/FINANCIETOREN
7ème étage / 7de verd. – bte/bus 71

Bd du Jardin Botanique / Kruidtuinlaan 50
B-1000 Bruxelles/Brussel

http://incc.fgov.be **http://nicc.fgov.be**