BNDMR
Banque Nationale de Données Maladies Rares

Cadre d'Interopérabilité Maladies Rares

CI-MR-1.1 : Spécifications IdMR

Auteur : BNDMR

Décembre 2014

Contenu

1 Objet

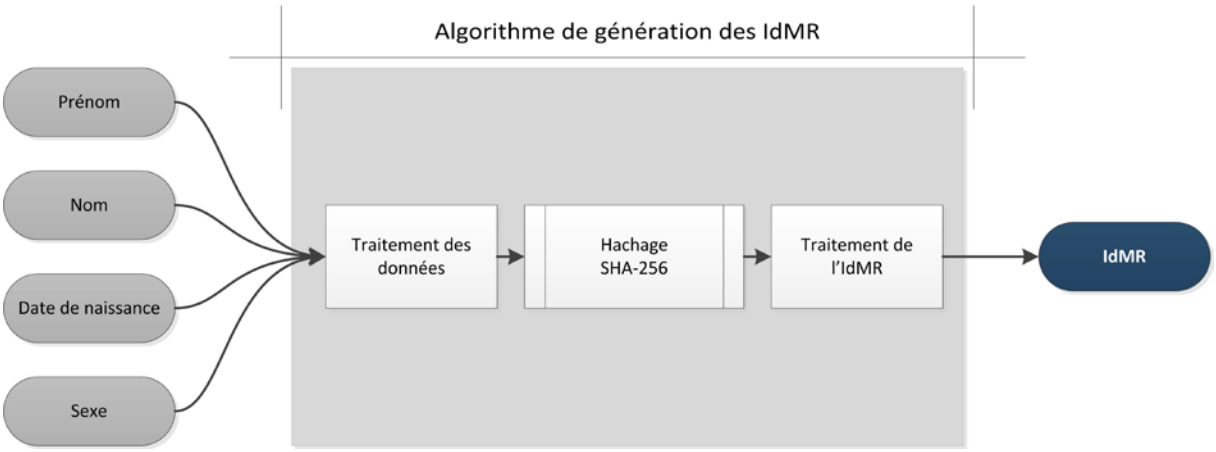
Le présent document vise à spécifier le processus de calcul de l'IdMR : l'identifiant utilisé au niveau de la Banque Nationale de Données Maladies Rares (BNDMR) afin d'identifier les patients maladies rares.

Il est destiné à tout organisme souhaitant générer cet identifiant afin de garantir une compatibilité au niveau de l'identification des patients lors d'échange de données avec la BNDMR.

2 Spécifications

2.1 Processus général

L'IdMR est une chaîne de 20 caractères numériques issue du hachage par la fonction SHA-256 de 4 éléments de données identifiants : le prénom, le nom, la date de naissance et le sexe du patient.



Les données en entrée de l'algorithme sont définies ci-dessous :

- *Prénom du patient* : prénom usuel du patient figurant parmi les prénoms inscrits sur l'acte de naissance.
- *Nom du patient* : patronyme, appelé aussi nom de famille ou nom de naissance du patient tel que déclaré sur l'acte de naissance.
- *Date de naissance du patient* : jour, mois et année de naissance.
- *Sexe du patient* : féminin, masculin ou indéterminé.

Remarque

L'IdMR ne sera pas calculé si au moins une de ces données est manquante.

Le cas particulier du fœtus :

- *Prénom : la lettre « f » pour fœtus suivie du numéro d'ordre dans la fratrie des fœtus dans le cas d'une grossesse gémellaire. Concaténer la chaine obtenue au prénom de la mère.*
  Exemple : f1Marta, f2Marta.
- *Nom : nom de naissance de la mère.*
- *Date de naissance : date de début de grossesse en ne gardant que le mois et l'année, le jour sera fixé au 1er jour du mois. Exemple : si la date de début de grossesse est estimé au 2014/11/11 la date utilisée pour le calcul de l'identifiant sera le 2014/11/01.*
- *Sexe : Par convention, pour le calcul de l'identifiant, le sexe sera fixé à « I » (inconnu) pour tous les fœtus.*

2.2 Traitement des données

2.2.1 Traitement des caractères

Les caractères acceptés sont seulement les caractères alphanumériques encodés UTF-8 (compatibilité encodage ASCII) : caractères alphabétiques de A à Z et les caractères numériques représentant les chiffres de 0 à 9.

Tout caractère accentué devra être remplacé par le caractère correspondant non accentué (*cf.* Annexe A).

Tout caractère spécial (symboles, espaces et ponctuation) sera supprimé.

Tout caractère alphabétique minuscule sera remplacé par le caractère majuscule correspondant.

2.2.2 Format des données

Prénom

Une fois le traitement des caractères effectué, le prénom sera tronqué pour ne pas dépasser une longueur de 10 caractères. Si sa longueur est inférieure à 10 caractères, il sera complété à droite par des espaces.

Une étude de la distribution de la longueur des prénoms et des noms pour 280 000 patients a été effectuée afin de déterminer le seuil qui sera appliqué à la longueur de ces données. Les médianes se situaient à 6.5 caractères pour le prénom et 7.1 caractères pour les noms. Le seuil de 10 caractères a été choisi parce qu'il permet de couvrir 75% de la population étudiée : c'est le troisième quartile.

Nom

Même traitement.

Date de naissance

La date de naissance est inscrite dans un format de 8 caractères numériques AAAAMMJJ (format ISO-8601 sans tirets).

Sexe

Utilisation du caractère alphabétique « F » si le patient est de sexe féminin, « M » si le patient est de sexe masculin ou « I » si le patient est de sexe indéterminé (ou inconnu pour les fœtus).

2.2.3 Concaténation

Les champs contenant les données traitées seront par la suite concaténés, sans l'ajout de séparateurs, suivant cet ordre : prénom, nom, date de naissance et sexe. Nous obtenons ainsi une chaîne primaire de 29 caractères.

Chaîne primaire (29) = prénom (10) + nom (10) + date de naissance (8) + sexe (1)

2.3 Hachage

La fonction de hachage à utiliser est le Secure Hash Algorithm SHA-256 défini dans la publication FIPS 180-2 par le National Institute of Standards and Technology (NIST) aux Etats-Unis et conforme en France au Référentiel Général de Sécurité[1] publié par l'Agence Nationale de la Sécurité des systèmes d'Information (ANSSI).

A cette étape du processus de génération de l'IdMR, la fonction de hachage transforme la chaîne primaire de 29 caractères en une empreinte de hachage de 256 bits.

2.4 Traitement de l'IdMR

L'empreinte de 256 bits est par la suite convertie en décimal. Chaque octet (8bits) des 32 octets en sortie du SHA-256 est « traduit » en un nombre décimal (d'une valeur allant de 0 à 255). Tous les zéros précédents ces nombres seront supprimés (25 au lieu de 025). Les 32 nombres ainsi obtenus sont concaténés en une chaîne de caractères. La chaîne est tronquée aux 20 premiers caractères (à gauche) : on obtient ainsi l'IdMR.

3 Risques et validation

3.1 Risques

Doublons :

Un patient se voit attribuer deux IdMR. Cela est dû aux modifications qui peuvent affecter les données en entrée : prénom, nom, date de naissance ou sexe. Les différentes raisons :

- Changement réel de l'une des données (changement du prénom par exemple). Ce cas est très rare mais il est recommandé de notifier l'équipe BNDMR de ce changement d'identifiant (envoi du couple ancien IdMR – nouveau IdMR) afin d'éviter la création d'un nouveau dossier patient dans la banque nationale.
- Erreur à la saisie des données. Pour éviter cela, il est conseillé d'intégrer des contrôles de qualité et de cohérence à la saisie et de signaler aux utilisateurs l'importance de l'exactitude de ces données.

Collisions :

Deux patients différents se voient attribuer un même IdMR. Ce risque est dû au caractère intrinsèque des fonctions de hachage puisqu'il existe moins d'empreintes possibles en sortie (taille 256 bits fixe) que de valeurs possibles en entrée. Par ailleurs, plus la chaîne de caractères retenue après le hachage est courte plus ce risque augmente (*cf.* 5.2).

[1] ANSSI (26 avril 2012) *Référentiel Général de Sécurité version 2.0, Annexe B1 - Mécanismes cryptographiques : Règles et recommandations concernant le choix et le dimensionnement des mécanismes cryptographiques.*

Autres risques :

Une mal interprétation des spécifications contenues dans le présent document ou une erreur lors de la programmation du générateur de l'IdMR conduirait à la production d'identifiants erronés. Afin d'éviter ce risque, un tableau de correspondances liant les données en entrée (personnes connues et décédées) à leurs idMR est joint à ce document (*cf.* annexe C). Il constitue une base de validation après la programmation du générateur de l'IdMR.

Des personnes malveillantes ayant accès aux dossiers patients anonymisés pourraient tenter de remonter à l'identité de ces patients. Pour ce faire, la reconstruction d'une table de correspondances liant toutes les valeurs possibles des données en entrée à leurs IdMR calculés pourrait constituer un recours. Cependant, il est à rappeler que les moyens qui devront être mis en place pour construire cette table sont très importants (avec un temps de calcul de l'ordre de plusieurs dizaines d'année).

3.2 Validation

Le processus de génération de l'IdMR a été validé après deux phases de test dont le but était de minimiser le risque de collisions :

Un test préliminaire a été fait sur les données nominatives de 45 000 personnes différentes. Avec un IdMR tronqué à 10 caractères, 6 collisions ont été détectées. Avec un IdMR tronqué à 20 caractères, aucune collision n'a été détectée.

Un deuxième test a été effectué sur les données nominatives de 280 402 patients. 2 458 doublons ont été détectés, il s'agissait de vrais doublons dans la base. Ainsi aucune collision n'a été générée pour les IdMR de 20 caractères.

Annexe A – Tableau de substitution des caractères

Caractère à substituer	Caractère substitut
À Á Â Ã Ä Å Æ à á â ã æ ä å	A
Ç ç	C
Ð ð	D
È É Ê Ë è é ê ë	E
Ì Í Î Ï ì í î ï	I
Ñ ñ	N
Ò Ó Ô Õ Ö Ø ò ó ô õ ö ø	O
Š š	S
Ù Ú Û Ü ù ú û ü	U
Ý Ÿ ý ÿ	Y
Ž ž	Z
Œ œ	OE
ß	SS
Caractères minuscules de ‘a’ à ‘z’	Caractères majuscules de ‘A’ à ‘Z’

Annexe B – Exemple de génération d'un IdMR

Phase de traitement des données :

	Avant traitement	Après traitement
Prénom	'Louis-René'	'LOUISRENE '
Nom	'des Forêts'	'DESFORETS '
Date de naissance	'1918-01-28'	'19180128'
Sexe	'M'	'M'

Phase de hachage et de post-traitement :

Chaîne primaire	'LOUISRENE DESFORETS 19180128M'
Hachage décimal	222 150 234 11 158 220 65 208 59 156 43 13 65 10 83 114 26 152 244 110 41 26 238 35 205 255 73 178 78 121 156 177
IdMR	22215023411158220652

Annexe C – Table de correspondances de validation

Prénom Nom Date de naissance Sexe	Chaîne primaire	IdMR
"Jean" "des Vallières" 1895-04-05 M	JEAN DESVALLIER18950405M	23112872142221771793
"Arthur" "Straußenburg" 1857-06-16 M	ARTHUR STRAUSSENB18570616M	52195118381273413616
"Louis-René" "des Forêts" 1918-01-28 M	LOUISRENE DESFORETS 19180128M	22215023411158220652
"Lucie" "Delarue-Mardrus" 1874-11-03 F	LUCIE DELARUEMAR18741103F	33163661851578420395
"Charles-Augustin" "Sainte-Beuve" 1804-12-23 M	CHARLESAUGSAINTEBEUV18041223M	23518514224810074791
"Victor" "Hugo" 1802-02-26 M	VICTOR HUGO 18020226M	21416852331492202521
"Alexandra" "David-Néel" 1868-10-24 F	ALEXANDRA DAVIDNEEL 18681024F	11871411851022441432
"François" "Nourissier" 1927-05-18 M	FRANCOIS NOURISSIER19270518M	16967145173172696162
"Eugène" "Labiche" 1815-05-06 M	EUGENE LABICHE 18150506M	22313519719914862056
"Jean-Jacques" "Ampère" 1800-08-12 M	JEANJACQUEAMPERE 18000812M	34218173806010193912