Documentation de l'interface d'alignement entre le référentiel des personnes physiques de la DJ et de la DDCOL (version 08/2020)

Table des matières

Objectif

Objectif général de l'alignement

L'objectif général est d'aligner deux jeux de données pour effectuer le rapprochement entre le système d'information juridique et celui documentaire. Ces jeux sont les suivants :

- les personnes physiques issues du système d'information juridique (DJ) :
  - un jeu avec les personnes ayant des contributions
  - un second jeu avec les personnes n'ayant pas de contributions
- les personnes physiques de la DDCOL

Données initiales

Côté DJ

Les données sont issues des tables PERSONNE, PARTENAIRE et CONTRIBUTION de la base ADAJE.

Matricule
Prénom
Nom
Pseudo nom
Pseudo prénom
Civilité
Date de décès
Commentaire homonymie
Contribution
Nombre de contributions

Table 1 Champs initialement disponibles côté DDCOL pour l'alignement

Côté DDCOL

Les données sont issues des tables concepts, conceptSchemes et texts_concepts du Lac de Données.

businessIdentifier
conceptIdentifier
Nom
Type de libellé
Note qualité
Sexe
Date de naissance
Date de décès
Type d'entité

Table 2 Champs initialement disponibles côté DDCOL pour l'alignement

Règles appliquées lors de l'alignement réalisé avec Talend

Après un premier traitement de normalisation des données de chaque côté (gestion des accents, de la casse, des particules, de la ponctuation, etc), deux jeux de données sont obtenus et seront utilisés pour la réalisation des alignements :

- un jeu unique pour la DJ

Matricule
Prénom
Nom
Pseudo nom
Pseudo prénom
Civilité
Date de décès
Commentaire homonymie
Contribution
Nombre de contributions

Table 3 Champs disponibles côté DDCOL pour les jointures

- un jeu pour la DDCOL

businessIdentifier
conceptIdentifier
NomConcept
Nom
Prénom
Pseudo
Type de libellé
Note qualité
Sexe
Date de naissance
Date de décès
Type d'entité

Table 4 Champs disponibles côté DDCOL pour les jointures

Priorisation de certains points de comparaison: calcul d'un indice de confiance

De multiples jointures sont possibles entre les jeux de données. Pour s'assurer une meilleure qualité des alignements produits, il est nécessaire de prioriser certains points de comparaison. Ainsi, les prénom et nom d'une personne sont prioritaires sur les pseudos (Voir : 4 étapes exclusives). De même, un alignement fait avec une comparaison supplémentaire sur la contribution/note qualité est plus sûr qu'un alignement réalisé sans (Voir : ci-dessous).

Pour évaluer la qualité des alignements réalisés, un compteur à points est mis en place pour chaque alignement. Ainsi, le barème suivant est appliqué :

Point de comparaison	Nombre de point(s)
Nom	1
Prénom	1
Pseudo_nom	1
Pseudo_prénom	1
Sexe	1
Date de décès	2
Contribution	2

Table 5 Concordance point de comparaison/nombre de points

Ce barème s'applique lors des jointures entre les jeux de données: chaque point de comparaison augmente l'indice de confiance de l'alignement. Ainsi, l'indice attribué aux alignements est compris entre 1 et 9.

4 étapes exclusives

L'attribution d'un indice de confiance n'étant pas suffisant (plusieurs matricules pouvant être alignés avec le même concept), plusieurs étapes, suivant une nouvelle fois une priorisation des points de comparaison, ont été réalisées :

1. jointures utilisant nom, prénom, pseudo-nom, pseudo-prénom et contributions
2. jointures utilisant nom, prénom, pseudo-nom et pseudo-prénom, sans correspondance des contributions
3. jointures sur les pseudos, et prénom de la DJ commençant par le prénom de la DDCOL (sauf pour les Jean et les Anne)
4. dernière jointure sur les nom et prénom seuls : cette jointure ayant un indice de confiance de 1, plusieurs matricules peuvent avoir le même concept. Cette jointure est considérée comme une aide à l'alignement manuel.

A l'issue de chaque étape, ce qui a été aligné suit un dernier traitement :

1. si plusieurs matricules ont été alignés avec le même concept, tous les alignements concernant ces matricules sont annulés;
   en revanche, un couple matricule/concept unique est considéré comme fiable et est conservé
2. une comparaison est effectuée sur les dates de décès :
   - si une date de décès est renseignée à la fois côté DJ et côté DDCOL et si elles correspondent, les points sont attribués et l'alignement confirmé ;
   - si une date de décès est renseignée à la fois côté DJ et côté DDCOL et si elles ne correspondent pas, l'alignement est annulé ;
   - si une date de décès est présente d'un côté mais pas de l'autre, l'alignement est confirmé mais ne prend pas les points correspondant à la date
3. les matricules et les concepts alignés sont retirés des jeux de données servant aux jointures qui suivront dans les autres étapes

Liste des jointures appliquées dans l'ordre

<table>
<tr><th>Côté DJ</th><th>Type de comparaison</th><th>Côté DDCOL</th><th>Indice de confiance</th><th>Volumétrie</th></tr>
<tr><td colspan="5">PREMIERE ETAPE : JOINTURES AVEC CONTRIBUTIONS</td></tr>
<tr><td colspan="3">JOINTURE N°1</td><td>6</td><td rowspan="6">9822</td></tr>
<tr><td>Nom</td><td rowspan="4">Egalité</td><td>Nom</td><td>1</td></tr>
<tr><td>Pseudo_nom</td><td rowspan="2">Prénom</td><td>1</td></tr>
<tr><td>Prénom</td><td rowspan="2">Prénom</td><td>1</td></tr>
<tr><td>Pseudo_prénom</td><td>1</td></tr>
<tr><td>Contribution</td><td>Contenue dans</td><td>Note qualité</td><td>2</td></tr>
<tr><td colspan="3">JOINTURE N°2</td><td>4</td><td>10861</td></tr>
</table>

Pseudo_prénom	Egalité	Prénom	1	
Pseudo_nom		Nom	1	
Contribution	Contenue dans	Note qualité	2	
JOINTURE N°3			**4**	**12445**
Nom	Egalité	Nom	1	
Prénom		Prénom	1	
Contribution	Contenue dans	Note qualité	2	
JOINTURE N°4			**4**	**10317**
Pseudo_nom	Egalité	Nom	1	
Pseudo_prénom		Prénom	1	
Contribution	Contenue dans	Note qualité	2	
A l'issue de cette étape, si des matricules différents sont alignés avec un même concept, les alignements concernés sont alignés. La date est ensuite vérifiée (**+2**). L'étape suivante est réalisée sans les 12330 matricules uniques et les 12330 concepts uniques alignés lors de cette étape.				**12330**
DEUXIEME ETAPE : JOINTURES SANS LES CONTRIBUTIONS				
JOINTURE N°5			**4**	**40587**
Nom	Egalité	Nom	1	
Pseudo_nom			1	
Prénom		Prénom	1	
Pseudo_prénom			1	
JOINTURE N°6			**2**	**48338**
Pseudo_nom	Egalité	Nom	1	
Pseudo_prénom		Prénom	1	
JOINTURE N°7			**2**	**58955**
Nom	Egalité	Nom	1	
Prénom		Prénom	1	
JOINTURE N°8			**2**	**45588**
Prénom	Egalité	Prénom	1	
Pseudo_nom		Nom	1	
A l'issue de cette étape, si des matricules différents sont alignés avec un même concept, les alignements concernés sont alignés. La date est ensuite vérifiée (**+2**).				**27416**

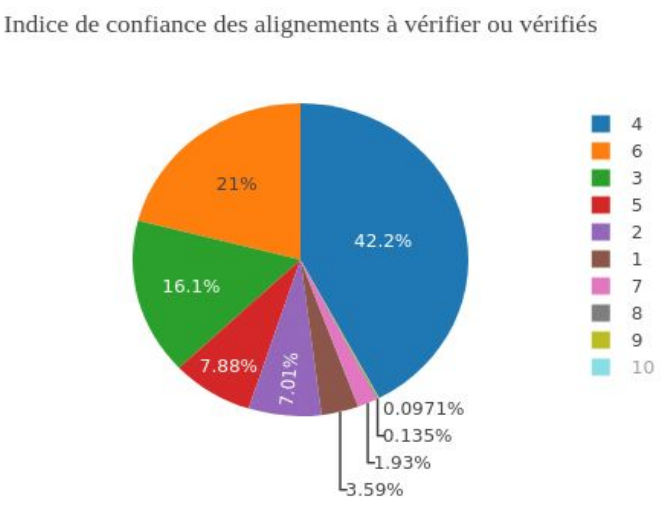
L'étape suivante est réalisée sans les 27416 matricules uniques et les 27416 concepts uniques alignés lors de cette étape.				
TROISIEME ETAPE : JOINTURES SUR LES PSEUDOS ET UTILISANT LES PRENOMS				
JOINTURE N°9			**3**	**624**
Pseudo_nom	Egalité	Pseudo	1	
Contribution	Contenue dans	Note qualité	2	
JOINTURE N°10			**1**	**7889**
Pseudo_nom	Egalité	Pseudo	1	
JOINTURE N°11			**2 (4)**	**493**
Nom	Egalité	Nom	1	
Prénom	Commence par	Prénom	1	
(Contribution)	(Contenue dans)	(Note qualité)	(2)	
JOINTURE N°13			**2 (4)**	**628**
Nom	Egalité	Nom	1	
Prénom	Existant des deux côtés mais différents	Prénom	1	
(Contribution)	(Contenue dans)	(Note qualité)	(2)	
A l'issue de cette étape, si des matricules différents sont alignés avec un même concept, les alignements concernés sont alignés. La date est ensuite vérifiée (**+2**). L'étape suivante est réalisée sans les 1953 matricules uniques et les 1953 concepts uniques alignés lors de cette étape.				**1953**
QUATRIEME ETAPE : NOM ET PRENOM SEULS				
JOINTURE N°13			**1**	**1494**
Nom	Egalité	Nom	1	
Prénom		Prénom		
				43193

Interface

Objectifs

Les objectifs de cette interface sont multiples:

- vérification des alignements réalisés avec Talend: la première étape de l'alignement réalisée avec Talend a permis d'aligner environ 60% des



matricules de la DJ avec des concepts de la DDCOL (les volumétries sont présentes sur la page d'accueil de l'interface). Puisque les alignements réalisés sont pour la plupart peu fiables en raison des homonymes et du peu d'éléments qui les ont permis, chaque alignement accompagné de son indice de confiance permettra d'aider la validation.

- création de nouveaux alignements: en raison des différences de graphie de l'état civil d'une même personne dans la DJ et dans la DDCOL, certains alignements n'ont pas pu être effectués. L'interface doit donc permettre de réaliser ces alignements grâce à la recherche du bon concept dans les données de la DDCOL.

Structure

Des alignements en base

Seuls les identifiants (matricule ou conceptIdentifier) sont concernés par les alignements. Les informations correspondant à ces identifiants (labels, dates, sexe, contributions, etc.) ne sont pas concernées par l'alignement puisque plusieurs labels peuvent correspondre à un même identifiant: la table des alignements n'est qu'une table de relation entre les deu jeux de données de la DJ et de la DDCOL; elle se compose des attributs suivants:

id_alignement
id_matricule
id_concept
indice_confiance
validation

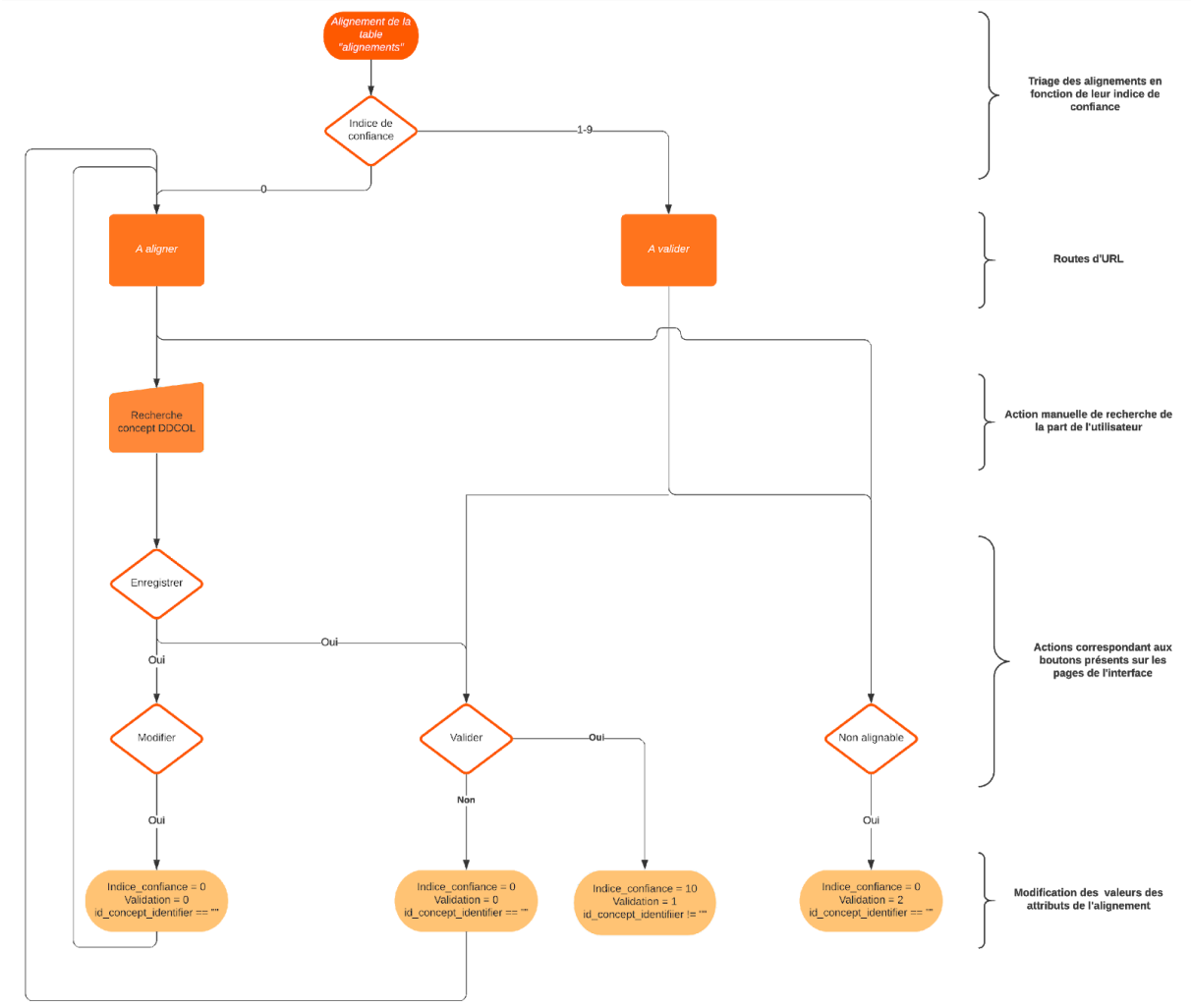
Table 5 Attributs de la table comportant les alignements

L'attribut "validation" n'est pas visible dans l'interface; il permet de connaître le statut de l'alignement de manière à orienter l'alignement dans le bon traitement dans l'interface:

- 0 est un alignement qui n'a été ni validé dans l'interface ni créé
- 1 est un alignement validé avec l'interface
- 2 est un alignement qu'il n'est pas possible d'aligner

De l'interface

Cette version de l'interface fonctionne sur une boucle infinie de modification et de rejet des alignements tant que l'alignement concerné n'est pas validé (valeur de l'attribut "validation" = 1) ou déclaré non alignable (valeur de l'attribut "validation" = 2) définitivement. Ainsi, un alignement avec 1 ou 2 en "validation" n'est plus modifiable de manière à supprimer cet alignement du circuit de validation ou de création d'alignements. Le logigramme suivant permet de visualiser le parcours d'un alignement dans l'interface:



Fonctionnalités

Plusieurs actions sont possibles sur un alignement de la table « alignements ». Un alignement peut présenter deux cas de figure :

— si l'indice de confiance est à 0 et la validation à 0, alors l'alignement n'a pas eu d'alignement proposé suite à l'étape Talend ; il faut donc l'aligner si possible

— si l'indice de confiance est compris entre 1 et 9, alors un alignement est proposé et il faut le valider ou le rejeter

Validation d'un alignement

La page « /alignements_a_valider » permet quatre actions :

— **Valider** l'alignement proposé : l'indice de confiance passe alors à 10, l'attribut validation à 1. Sur la page du matricule (« /matricule/<id_matricule> »), plus aucune action n'est alors possible ; l'alignement est considéré comme définitif et sûr.

— **Rejeter** l'alignement proposé: l'indice de confiance passe à 0, l'attribut validation reste à 0, et l'« id_concept » qui était présent est supprimé. Cet alignement rejeté sera alors visible dans la page « /alignements_a_faire ».

— Alignement **non alignable** : l'indice de confiance passe à 0, et la validation à 2. Sur la page du matricule (« /matricule/<id_matricule> »), plus aucune action n'est alors possible ; l'alignement impossible est considéré comme définitif et sûr.

— **Suivant** : permet de se diriger vers un alignement suivant. Aucune action n'est alors effectuée sur l'alignement actuellement proposé.

Alignement à réaliser

La page « /alignements_a_faire » propose deux actions quand une recherche de concept DDCOL n'a pas encore été effectuée :

— Alignement **non alignable** : l'indice de confiance passe à 0, et la validation à 2. Sur la page du matricule (« /matricule/<id_matricule> »), plus aucune action n'est alors possible ; l'alignement impossible est considéré comme définitif et sûr.

— **Suivant** : permet de se diriger vers un alignement suivant. Aucune action n'est alors effectuée sur l'alignement actuellement proposé.

Pour aligner le matricule DJ avec un concept DDCOL, il faut effectuer une recherche manuelle selon trois critères (ne remplir qu'un seul est suffisant) :

Figure Barre de recherche dans les concepts de la DDCOL

— le champ « nom » recherche exactement le nom donné dans la DDCOL. La recherche n'est pas sensible à la casse et à l'accentuation.

— le champ « prénom » recherche exactement le prénom donné dans la DDCOL. La recherche n'est pas sensible à la casse et à l'accentuation.

— la chaîne rentrée dans le champ « état civil libre » est recherchée dans la DDCOL. La recherche n'est pas sensible à la casse et à l'accentuation.

Si des résultats apparaissent, il est possible de cocher le concept avec lequel on souhaite aligner le matricule. Un bouton « **Enregistrer** » disponible en haut de page permet d'enregistrer ce choix : l'indice de confiance passe à 5, la validation reste à 0, et l'id du concept choisi est inséré dans « id_concept ». Une redirection est effectuée vers la page du matricule « /matricule/<id_alignement>».

Sur cette page de redirection, plusieurs actions sont désormais possibles :

— **Valider** l'alignement proposé : l'indice de confiance passe alors à 10, l'attribut validation à 1. Sur la page du matricule (« /matricule/<id_matricule> »), plus aucune action n'est alors possible ; l'alignement est considéré comme définitif et sûr.

— **Rejeter** l'alignement proposé: l'indice de confiance passe à 0, l'attribut validation reste à 0, et l'« id_concept » qui était présent est supprimé. Cet alignement rejeté sera alors visible dans la page « /alignements_a_faire ».

— **Modifier** l'alignement proposé : l'indice de confiance passe à 0, l'attribut validation reste à 0, et l'« id_concept » qui était présent est supprimé. Cet alignement modifié sera alors visible dans la page « /alignements_a_faire ».

— **Suivant** : permet de se diriger vers un alignement suivant. Aucune action n’est alors effectuée sur l’alignement actuellement proposé.