Documentation de l'alignement des notes qualités de la DDCOL avec le thésaurus des noms communs

Objectifs

Objectifs

Plusieurs objectifs sont à l'origine de l'alignement des fonctions des notes qualité contenues dans la description des personnes physiques et morales de la DDCOL avec le thésaurus de noms communs:

- extraire les fonctions des notes qualités rédigées en texte libre
- rassembler ces fonctions quand elles sont identiques
- aligner ces fonctions avec leur équivalent dans le thésaurus de noms communs afin d'obtenir le conceptIdentifier du terme du thésaurus présent dans le Lac de données
- classer ces fonctions selon les catégories du thésaurus
- extraire les fonctions non présentes dans le thésaurus afin de créer les entrées ou de corriger les notes qualités

Jeux de données

Les notes qualités se trouvent dans le référentiel des personnes physiques et morales (PP/PM) de la DDCOL. Elles font l'objet d'un attribut d'une table de la base de données et sont rédigées en texte libre. Environ 400000 personnes possèdent ces notes qualités.
Ces notes qualités sont rédigées suivant des règles écrites dans une documentation disponible lors du catalogage au dépôt légal. Deux types de notes qualités se distinguent par la norme qui dirige leur création:

- un premier type, le plus courant, a la forme suivante:
  Fonction1, fonction2, ... Pays1, pays2, ...
- le second type, plus rare, est composé ainsi:
  Homonymes: 1- Fonction1, fonction2, ... Pays1, pays2, ... ; 2 - Fonction1, fonction2, ... Pays1, pays2, ... ; ...

Ces notes qualités reposent sur une ponctuation explicite qui permet à elle seule l'extraction des fonctions: la virgule sépare deux entités du même type, tandis que le point sépare les fonctions des pays. Le point virgule, dans le cas des homonymes, sépare ces derniers en complément du chiffre et du tiret.
La rédaction de ces notes étant réalisée par un agent humain, l'oubli ou la mauvaise utilisation d'une ponctuation conduit à des difficultés pour l'extraction des fonctions et des pays.

Le thésaurus de noms communs utilisé pour cet alignement utilise les termes dont la facette est "$Personne", c'est à dire que le terme se rapporte à une personne. Plusieurs informations sont disponibles pour chacun des termes:

- le **terme préférentiel** de l'entrée de thésaurus: ce terme est celui qui est aligné avec les fonctions des notes qualités car ils ont une forme identique et se rapportent à un métier (par exemple, le thésaurus a le terme "chef cuisinier" plutôt que "cuisine", ce qui correspond au "chef cuisinier" des notes qualités)
- l'**arborescence** de ce terme préférentiel: cette arborescence n'est pas exploitable car ses termes ne sont pas des métiers mais des termes génériques ("sportif" n'est pas un terme du thésaurus, alors qu'il existe un terme ascendant "sport": les notes qualités avec "sportif" n'auront pas d'alignement)
- le **terme directement parent** du terme préférentiel, issu de cette arborescence
- les **synonymes** du terme préférentiel: ils sont utilisés dans l'alignement car ils se rapportent à un métier
- la **facette**, qui est toujours "$Personne"
- une **précision** sur l'ascendance (le parent du parent), issue de l'arborescence
- la **catégorie de rattachement** du terme: cette catégorie est reprise pour la classification finale

Traitements préalables

La plupart des traitements préalables concerne les notes qualités. En effet, les termes du thésaurus sont déjà des termes contrôlés: les fonctions des notes qualités doivent, par conséquent, devenir des termes contrôlés construits sur la même forme que les termes du thésaurus.

D'abord, il est nécessaire d'écarter de l'alignement les notes qualités ne comportant pas de fonctions, mais seulement des indications reliant la personne à un événement ou à un fait judiciaire: les notes qualités commençant par "Attentat", "Scandale", "Faits divers", "Meurtre", "Victime", ... sont écartées. De même, les notes qualités n'indiquant que la filiation ou l'appartenance à une famille ne sont pas traitées.

La première étape est la **scission** des notes qualités qui sont de type "Homonymes": en les scindant, on obtient plusieurs notes qualités, dépourvues du mot "Homonymes", du double point, des chiffres et des tirets.

Ensuite a lieu la **scission** des notes qualités sur le point, afin de séparer les fonctions et les pays. À partir de cette étape, seules les fonctions sont désormais disponibles dans le jeu de données, tout ce qui n'a pas d'utilité dans l'alignement ou le futur classement est supprimé.

La troisième étape consiste en une nouvelle **scission**, celle des fonctions entre elles. En effet, certaines personnes ont jusqu'à huit fonctions indiquées dans leur note qualité. Cette scission a lieu sur la virgule qui sépare ces fonctions. À l'issue de cette étape, une liste de fonctions est obtenue:

une ligne est une seule et unique fonction. Cependant, des noms de pays de pays se trouvent dans cette liste en raison d'une mauvaise utilisation de la ponctuation lors du catalogage: une virgule a été mise à la place du point pour séparer les fonctions et les pays.

Une quatrième étape permet de **normaliser** l'ensemble des fonctions pour correspondre au maximum aux termes du thésaurus:

- les majuscules sont transformées en minuscules
- les pluriels sont supprimés pour avoir un ensemble de termes au singulier
- les féminins sont transformés en masculins
- la ponctuation est supprimée
- l'accentuation est également enlevée

Cette quatrième étape est aussi réalisée sur les termes et les synonymes du thésaurus: l'application des mêmes règles de chaque côté doit permettre des alignements facilités. Dans le thésaurus, les modifications de graphie seront moins nombreuses que dans les fonctions extraites des notes qualités, mais elles sont tout de même nécessaires.

À l'issue de ces étapes préparatoires à l'alignement, la liste de fonctions normalisées extraites des notes qualités montrent de très nombreuses coquilles et erreurs humaines dans la rédaction des notes qualité: la fonction de "politique" (initialement "homme-" ou "femme politique") comporte plusieurs dizaines de variantes.

L'alignement

De manière à contourner les différences de graphie, ainsi que les différences dans la désignation d'une même fonction, plusieurs jointures sont effectuées entre les jeux de données[1]:

- une première a lieu sur la stricte égalité des termes (400 alignements réalisés ainsi sur les 37000 fonctions distinctes extraites des notes qualités)
- ensuite, certaines fonctions ne peuvent pas être alignées en raison d'un terme polysémique et générique placé en première position, comme "chef" ou bien "maître". Afin d'aligner le thésaurus avec le second terme, la RegEx suivante est utilisée pour ne pas tenir compte des premiers termes:
  ^(directeur|directrice|maitre|presidente?|chef|assistante?|experte?|specialiste|fondateur|co[
  |-]?fondateur|createur|creatrice|concepteur|conceptrice|inventeur|organisateur|organisatrice|collaborateur|porte[
  |-]parole|participant|docteur|responsable|membre|passionnee?|designer|ancien|ex|co[
  |-]?president|sous|vice|concepteur|conceptrice|co|haute?|joueur|organisateur)s?
  La jointure a lieu ensuite sur la stricte égalité des seconds termes des fonctions avec le thésaurus (95 alignements).
- la précédente étape n'ayant retourné que peu d'alignements, une nouvelle jointure sur les deuxièmes termes est nécessaire, non plus avec une stricte égalité des termes de chaque côté, mais avec une fonction des notes qualités qui commence par le terme du thésaurus (900 alignements supplémentaires sont ainsi réalisés).
- enfin, la dernière étape consiste en faisant une jointure similaire sur les premiers termes des fonctions (17700 alignements)

[1] Après chaque étape, ce qui n'a pas été aligné sert de jeu de données à l'étape suivante.

À l'issue de ces jointures, ce sont environ 50% des fonctions des notes qualités qui ont pu être alignées. Les fonctions qui ne l'ont pas été n'ont soit pas d'équivalent dans le thésaurus en raison d'une trop grande spécificité des termes, soit pas les mêmes graphies (la grande majorité de ces cas n'est pas alignée en raison de coquilles).

Cependant, un nombre assez important de fonctions a tout de même un équivalent dans le thésaurus, mais n'est pas aligné. En effet, l'extraction du thésaurus de noms communs a été faite sur les facettes "$PERSONNE". Les fonctions relatives aux professions de l'audiovisuel sont néanmoins présentes dans la facette "PERSONNE-AUDIOVISUEL". Cette nouvelle facette, ainsi que d'autres qui ne sont pas citées ici, est par conséquent à ajouter aux futurs lancements des jobs Talend qui seront faits.

Le classement

Le premier objectif de l'alignement du thésaurus de noms communs et des notes qualités était d'obtenir l'identifiant du concept correspondant aux fonctions extraites des notes qualités. Dans un second temps, il s'agit aussi d'assurer la qualité des notes qualités et du thésaurus afin d'associer à chaque note qualité le maximum de concepts du thésaurus.

Pour assurer le suivi de cette qualité, il est nécessaire:

- de compter le nombre d'occurrences de chaque fonction dans l'ensemble des fonctions extraites des notes qualités afin de privilégier la modification des fonctions présentes de nombreuses fois dans les notes qualités
- de trier les fonctions des notes qualités selon quelques thématiques pour en connaître l'étendue
- de connaître la provenance de chaque classement

Tout d'abord, il convient de connaître le nombre de personnes physiques ou morales contenant une même fonction. Ce **comptage** est aisé à réaliser et permet de remarquer une grande disparité dans la description des fonds: des termes communs comme "Réalisateur" ou "Homme politique" sont présents des dizaines de milliers de fois, alors que pour des fonctions plus spécifiques la dénomination varie (par exemple, "joueur de basket", joueur de basketball").

Ensuite, un **classement** de ces fonctions doit être effectué: la réflexion avec le métier a conduit à l'adoption des huit catégories existantes dans le thésaurus, ajoutées à une neuvième catégorie "RESTE" et une dixième "RESTE_FAITS_DIVERS" issue des notes qualités écartées au début du traitement:

- Art et culture
- Communication, diffusion, traitement de l'information
- Sciences
- Sciences humaines
- Sport
- Vie économique
- Vie quotidienne, habitat, alimentation et loisirs
- Vie sociale

L'attribution de ces catégories se fait par différents moyens (chaque moyen utilisé est indiqué dans les données finales dans la colonne "source_alignement" afin de connaître la provenance du classement de chaque terme[2]):

- “NC” est l’indication d’une fonction classée grâce à l’alignement effectué précédemment et à la catégorie qui était attachée au terme du thésaurus: “NC” ne concerne par conséquent que les 50% d’alignements réussis
- “deuxième_terme” concerne uniquement les alignements sur les seconds termes des notes qualités avec les termes du thésaurus. Comme pour les “NC”, la catégorie provient de la catégorisation attachée au terme du thésaurus.
- “aide_DL” indique que la fonction issue des notes qualités n’a pas été alignée avec le thésaurus et qu’elle a été classée selon les règles de catalogage en vigueur au Dépôt Légal (voir capture d’écran ci-dessous). Des domaines découpent le manuel de rédaction des notes qualité: ces domaines permettent, pour chaque fonction associée, de définir la catégorie de classement de chaque fonction extraite des notes qualités.

Arts et artisanat	Acteur (actrice) et non comédien, tragédien ou autre terme	Troupe de théâtre
	Acteur-humoriste	
	Chanteur (chanteuse) + genre *exemples* : Chanteur variétés Chanteur rock Chanteur jazz	**Groupe musical ou Groupe + spécialité** **Noter également vocal ou instrumental** *exemples* : **Groupe rock** **Groupe hip hop** **Groupe vocal traditionnel de musique celte**
	Chanteur lyrique (chanteuse) + voix *exemple* : Chanteur lyrique baryton	
	Chef d'orchestre	Orchestre ou Ensemble + précision (jazz, classique, baroque, de chambre etc) *exemples* : Orchestre de jazz Orchestre classique Ensemble baroque

- “catégorisation_supplémentaire”. À l’issue des trois premiers classements, il a été remarqué l’absence de catégorisation pour certaines fonctions très souvent employées comme “enseignant”. Cette catégorisation ne repose sur aucune règle: elle est effectuée *à la main* dans Talend pour toutes les fonctions de plus de 100 occurrences qui n’ont pas été classées.

Conclusion

L’absence d’alignement puis de classement est dû à plusieurs facteurs:

- une extraction des noms communs du thésaurus incomplète en raison d’une migration pas encore effectuée dans le Lac de données
- des fonctions extraites des notes qualités n’en sont pas, comme “cirque pinder”

[2] La connaissance de cette provenance est indispensable pour pouvoir évaluer les opérations à effectuer ensuite: si la provenance est “Deuxième terme”, alors aucune opération sur la note qualité n’est nécessaire; en revanche, si la provenance est “Catégorisation supplémentaire”, il sera nécessaire soit d’ajouter la fonction au thésaurus, au bien de corriger la note qualité pour que l’alignement se fasse.

- des fonctions trop spécifiques ne sont pas présentes dans le thésaurus, comme “chemisier”
- les coquilles dans la rédaction des notes qualité crée beaucoup de fonctions qui n’ont qu’une occurrence et qui ne sont pas alignables