Documentation des alignements de personnes, séries et fictions avec Wikidata

Objectif

Objectif général

L'objectif de ces trois alignements, réalisés distinctement, est de récupérer l'identifiant Wikidata de chaque instance ou concept. Pour les personnes physiques, les identifiants VIAF, BnF et ISNI sont également à obtenir. Pour les épisodes de séries et les fictions, il s'agit également de récupérer l'ISAN . Enfin, dans le cas de l'alignement des épisodes de série, il s'agit d'obtenir l'identifiant Wikidata de la série à laquelle l'épisode appartient, ainsi que celui de l'épisode lui-même.

Jeux de données

Personnes physiques et morales

Les concepts des personnes utilisés dans l'alignement proviennent du référentiel des Personnes physiques et morales de la DDCOL. Les attributs de nom, de sexe, de dates de naissance et de décès, et la note qualité sont disponibles pour effectuer cet alignement.

Tous les concepts ne sont pas concernés par l'alignement avec Wikidata: certains ont déjà leur identifiant Wikidata stocké dans le Lac, d'autres qui sont des personnes morales ne sont pas pris en compte. Ce sont 340000 concepts environ qui nécessitent la recherche de leur identifiant Wikidata.

Séries

Les séries à aligner proviennent du Lac de données: ces instances n'ont pas encore toutes d'ISAN, et aucune n'est liée à Wikidata. Les titres d'épisode et de série sont disponibles, ainsi que l'année de production, le genre de la série, et les numéros d'épisode et de saison. Ce sont environ 176000 épisodes qui sont à traiter.

Fictions

De même que les séries, les fictions sont extraites du Lac. Moins nombreuses (40000 environ), elles disposent d'un titre, d'un réalisateur et d'un producteur, d'une année de producteur et d'un genre.

Traitements préalables

L'ensemble des alignements nécessite un traitement préalable des données présentes à l'INA avant de pouvoir chercher l'entité de Wikidata qui leur correspond. Ce traitement préalable concerne notamment la graphie des titres ou des noms de personnes:

- les noms de personnes sont sous la forme *Nom, Prénom* alors que les libellés (préférentiels ou alternatifs) de Wikidata sont sous la forme *Prénom Nom*: l'ensemble des noms et prénoms de la DDCOL sont inversés pour obtenir la forme de Wikidata
- le pays de nationalité d'une personne est extrait de la note qualité à partir du point caractéristique qui sépare la fonction du pays; par conséquent, la fonction de la personne est en même temps extraite
- toutes les dates sont rapportées à la seule année
- les accentuations et les ponctuations sont supprimées

Ces traitements, s'ils concernent d'abord les données de l'INA, seront également appliqués aux données extraites de Wikidata avant de procéder aux différents alignements: les jeux de données à aligner doivent respecter les mêmes règles afin de concorder au maximum lors de la réalisation de l'alignement.

Requêtes dans Wikidata

Deux méthodes d'accès aux entités Wikidata et à leurs déclarations sont disponibles. L'accès par le Sparql-EndPoint est adapté aux requêtes uniques, non répétées automatiquement, et n'étant pas trop lourdes de manière à ne pas renvoyer de timeout. Si les requêtes à effectuer sont répétitives, résultent d'une itération sur un jeu de données, ou très nombreuses, l'API Wikibase doit être utilisée: elle n'a aucune limite de nombre de requêtes par minute et offre de multiples modules pour adapter les requêtes.

Le **Sparql-EndPoint** est utilisé en construisant des URLs avec deux paramètres:

- *q*: la requête Sparql encodée est placée après ce paramètre
- *format*: le format de sortie doit être indiqué; il peut être JSON ou XML

Cet outil de récupération des données de Wikidata est utilisé dans le cas de requêtes précises et uniques, non répétées automatiquement. Dès qu'il y a utilisation de plusieurs options dans la requête, ou utilisation de l'ontologie wikibase pour obtenir des libellés, un timeout apparaît et rend impossible la récupération des données.

L'**API Wikibase** (https://wikidata.org/w/api.php) est aussi accessible par une URL qu'il convient de construire selon des modules et des paramètres. Deux paramètres sont constamment utilisés:

- *action:* ce paramètre permet d'indiquer le module utilisé
- *format*: il s'agit de l'indication du format de sortie (nous choisissons le JSON comme pour le Sparql-EndPoint)

Les autres paramètres d'URL sont variables en fonction des modules utilisés. Pour ces alignements, nous avons utilisé les modules et les paramètres associés suivants:

- *wbsearchentities*: ce module permet la recherche d'une chaîne de caractères dans les libellés des entités de Wikidata; les paramètres d'URL suivants sont par conséquent utilisés:
    - *search* pour indiquer la chaîne de caractères à rechercher
    - *language* pour spécifier la langue du libellé dans lequel rechercher
    - *limit* pour indiquer le nombre de résultats maximum souhaités
- *wbgetentities*: il permet d'obtenir les données associées aux entités recherchées
    - *ids* spécifie jusqu'à cinquante identifiants d'entités dont on souhaite obtenir les données liées
    - *props* permet d'indiquer quel type de données on souhaite en retour: nous utilisons principalement *claims, labels, aliases*
    - *languages* indique l'ensemble des langues des textes obtenus en retour
- *wbgetclaim*s: ce module est une spécification de wbgetentities. Il permet, pour un seul identifiant d'entité, d'obtenir l'ensemble des déclarations associées
    - *entity* spécifie l'identifiant de l'entité recherchée

L'API Wikibase offre la possibilité de pouvoir effectuer plusieurs requêtes simultanément puisque l'adresse IP du client n'est pas vérifiée. De plus, l'obtention des résultats est beaucoup plus rapide que celle du Sparql-EndPoint.

Réalisation des alignements

Alignement des personnes physiques

Deux étapes constituent cet alignement afin de tenir compte des données présentes dans la DDCOL.

La **première étape** utilise les types données dont la propriété Wikidata est facilement identifiable et dont la valeur sera proche de celle de la DDCOL. Ainsi, cette étape n'utilise pas les fonctions, mais les autres données disponibles pour chaque personne:

- le sexe de la personne est obtenu à partir de la propriété P21
- la date de naissance à partir de P569
- la date de décès à partir de P570
- le pays d'exercice des fonctions est obtenu à partir de P27

Grâce au traitement préalable des données de la DDCOL, ces quatres informations disposent de données comparables à celles de Wikidata. Pour cela, il faut d'abord aligner le sexe et le pays avec leurs entités Wikidata: l'obtention des pays se fait avec l'URL indiquée en note[1] (les libellés

[1] https://query.wikidata.org/sparql?query=SELECT%20DISTINCT%20%3Fcountry%20%3FcountryLabel%20%3FaltLabel%0AWHERE%0A%7B%0A%20%20%3Fcountry%20wdt%3AP31%20wd%3AQ3624078%20.%0A%20%20%23not%20a%20former%20country%0A%20%20FILTER%20NOT%20EXISTS%20%7B%3Fcountry%20wdt%3AP31%

préférentiels et alternatifs des entités récupérées sont alignés avec les pays extraits des notes qualités).
Après la création de ces quatre points de comparaison avec Wikidata, il est nécessaire de comparer les personnes de la DDCOL et les entités de type Humain de Wikidata. Pour cela, une requête Sparql par personne est possible dans le Sparql-EndPoint. Seulement, la spécification de quatre points de comparaison, et l'utilisation de l'ontologie wikibase pour récupérer les libellés en français, entraîne un temps de réponse compris en 20 et 40 secondes. Ainsi, pour traiter les 340000 concepts, il faudrait une dizaine de jours (2800 heures environ).
Pour contourner ces performances faibles, l'utilisation de l'API est essentielle:

- d'abord, le module *wbsearchentities* est utilisé afin de récupérer, pour chaque concept, les entités dont les libellés s'approchent du nom du concept: plusieurs entités peuvent par conséquent être retournées comme résultats, notamment dans le cas des homonymes
- ensuite, avec le module *wbgetclaims*, pour chaque entité obtenue, les valeurs des déclarations des propriétés P569, P570, P27 et P21 sont stockées
- enfin, le concept de la DDCOL est comparé à chaque entité selon les valeurs des déclarations obtenues ci-dessus

Cette première étape ne suffit pas pour aligner l'ensemble des concepts, notamment en raison des lacunes dans les données de la DDCOL (absence de dates, absence de pays) qui n'offrent par conséquent pas assez de points de comparaison.

La **seconde étape** s'appuie sur les fonctions des notes qualités. Ces fonctions, comme les libellés qui seront récupérés sur Wikidata, feront l'objet d'une normalisation comme indiqué en seconde partie. La propriété P106 accueille la mention de métier dans les déclarations des entités. Wikidata étant un graphe, il est nécessaire de trouver l'entité qui contient l'ensemble des fonctions professionnelles afin de facilement pouvoir récupérer les entités et leurs déclarations.

Trois premières requêtes sont nécessaires afin de récupérer seulement les **identifiants et les libellés des métiers** et des fonctions présents dans Wikidata. Leur nombre étant très élevé, le Sparql-EndPoint renvoie une erreur, ce qui contraint à la réalisation de ces trois requêtes successives:

- L'entité Q61788060 (Activités humaines) comprend l'ensemble des professions et des fonctions professionnelles. Une première requête est effectuée dans le Sparql-EndPoint[2] afin de récupérer les entités directement liées aux Activités Humaines (i.e. qui ont une propriété P31 "est une instance de").
- Une autre requête Sparql est effectuée pour récupérer les sous-classes directes des Activités Humaines (avec la propriété P279 "est une sous-classe de ")[3]

20wd%3AQ3024240%7D%0A%20%20%23and%20no%20an%20ancient%20civilisation%20%28needed%20to%20exclude%20ancient%20Egypt%29%0A%20%20FILTER%20NOT%20EXISTS%20%7B%3Fcountry%20wdt%3AP31%20wd%3AQ28171280%7D%0A%20%20OPTIONAL%7B%3Fcountry%20skos%3AaltLabel%20%3FaltLabel%20.%20FILTER%20%28lang%28%3FaltLabel%29%20%3D%20%22fr%22%29%7D%20.%0A%0A%20%20SERVICE%20wikibase%3Alabel%20%7B%20bd%3AserviceParam%20wikibase%3Alanguage%20%22fr%22%20%7D%0A%7D%0AORDER%20BY%20%3FcountryLabel&format=json

[2] https://query.wikidata.org/sparql?query=select%20%3Ff%20%3FfLabel%20%3FaltLabel%0Awhere%7B%0A%3Ff%20wdt%3AP31%20wd%3AQ61788060.%0A%20%20OPTIONAL%7B%3Ff%20skos%3AaltLabel%20%3FaltLabel%20FILTER%20%28LANG%28%3FaltLabel%29%20%3D%20%22fr%22%29%7D%0A%20%20SERVICE%20wikibase%3Alabel%20%7B%20bd%3AserviceParam%20wikibase%3Alanguage%20%22fr%22.%20%7D%0A%7D&format=json

- Enfin, la troisième est réalisée à partir d'une itération sur les sous-classes récupérées précédemment afin d'obtenir leurs sous-classes et toute leur descendance avec P279* (ce qui était impossible à réaliser dans les deux premières requêtes à cause du Timeout du Sparql-EndPoint)[4]

À l'issue de ces trois étapes, l'ensemble des entités liées à Activités Humaines est stocké en local avec les identifiants et les libellés préférentiels et alternatifs. Ainsi, les fonctions extraites des notes qualités peuvent être alignées avec leur entité de Wikidata selon les libellés.

Cependant, la fonction associée à un concept de l'INA et celle associée à une entité de Wikidata ne sont pas nécessairement les mêmes: l'une peut être plus spécifique que l'autre, ou inversement. Pour contourner ce problème, il est nécessaire d'utiliser la hiérarchie de Wikidata faite avec les propriétés P31 et P279: en considérant la classe-mère de la fonction indiquée en P106, ainsi que les sous-classes associées à cette fonction avec la propriété P279, le champ de comparaison est élargi et permet le plus souvent d'aligner le concept grâce à cette classe-mère ou aux sous-classes.

Cet alignement de personnes physiques de la DDCOL avec Wikidata a un rendement de 10%: 30000 concepts ont pu être alignés, et leurs identifiants VIAF, BNF et ISNI récupérés en plus de l'identifiant Wikidata. Les raisons de ce faible nombre d'alignements sont multiples:

- les noms du concept et de l'entité sont différents, surtout en cas de noms d'origine étrangère
- les fonctions ne correspondent pas, ou bien leurs libellés ne sont pas identiques à ceux des fonctions des concepts
- la personne du concept est absente de Wikidata car elle n'est pas connue

Alignement des épisodes de séries et des fictions

L'alignement des épisodes de séries et des fictions fonctionne selon la même méthode que les personnes physiques: il est d'abord nécessaire de récupérer localement les identifiants des entités de classe Série Télévisée (Q5398426) ou Film (Q11424), puis leurs libellés et quelques déclarations.

D'abord, les **identifiants et les libellés** français et anglais (les séries et les films ne sont généralement pas tous traduits) des **sous-classes** de Série Télévisée et Fictions sont récupérés avec le Sparql-EndPoint[5].

---

3 https://query.wikidata.org/sparql?query=select%20%3Ff%20%3FfLabel%20%3FaltLabel%0Awhere%7B%0A%3Ff%20wdt%3AP279%20wd%3AQ61788060.%0A%20%20OPTIONAL%7B%3Ff%20skos%3AaltLabel%20%3FaltLabel%20FILTER%20%28LANG%28%3FaltLabel%29%20%3D%20%22fr%22%29%7D%0A%20%20SERVICE%20wikibase%3Alabel%20%7B%20bd%3AserviceParam%20wikibase%3Alanguage%20%22fr%22.%20%7D%0A%7D&format=json

4 https://query.wikidata.org/sparql?query=select%20%3Ff%20%3FfLabel%20%3FaltLabel%0Awhere%7B%0A%3Ff%20wdt%3AP279%20wd%3A + identifiant de la sous-classe + .%0A%20%20OPTIONAL%7B%3Ff%20skos%3AaltLabel%20%3FaltLabel%20FILTER%20%28LANG%28%3FaltLabel%29%20%3D%20%22fr%22%29%7D%0A%20%20SERVICE%20wikibase%3Alabel%20%7B%20bd%3AserviceParam%20wikibase%3Alanguage%20%22fr%22.%20%7D%0A%7D&format=json

5 Pour les séries télévisées: https://query.wikidata.org/#select%20%3Fserie%20%3FserieLabel%0Awhere%7B%0A%3Fserie%20wdt%3AP279%2a%20wd%3AQ5398426.%0Aservice%20wikibase%3Alabel%7Bbd%3AserviceParam%20wikibase%3Alanguage%20%22fr%2Cen%22%7D%0A%7D

Ensuite, les **identifiants des instances** de chacune des sous-classes obtenues peuvent être recherchés. Sparql est une nouvelle fois utilisé: le but n'est pas de récupérer les déclarations des entités, seulement les identifiants des entités qui sont des instances des sous-classes de Séries Télévisées ou de Fictions[6].

À l'issue de ces requêtes Sparql, les identifiants des entités sous-classes de Séries Télévisées ou de Fictions sont stockés localement. Il est désormais nécessaire de récupérer les **déclarations** dont les valeurs pourront être comparées aux données de l'INA. L'API Wikibase permet cette récupération, effectuée par lots de cinquante comme le permet l'URL, avec le module *wbgetentities*.

Dans le cas des **fictions**, les valeurs des propriétés suivantes sont récupérées:

- P577 (date de publication)
- P57 (réalisateur)
- P162 (producteur)
- libellés préférentiels et alternatifs du titre et des noms de personnes (réalisateur et producteur)

Dans le cas des épisodes de **séries télévisées**:

- P577 (date de publication)
- P179 (série d'appartenance de l'épisode)
- P4908 (nom de la saison)
- qualificatif P1545 de P179 (numéro d'épisode)
- qualificatif P1545 de P4908 (numéro de saison)
- libellés préférentiels et alternatifs en français et en anglais des titres
- libellés préférentiels et alternatifs en français des valeurs des propriétés

La récupération de ces données de Wikidata permet d'aligner les données de l'INA: 50% des fictions ont pu être alignées, tandis que 10% des épisodes de séries sont alignés. Cependant, de même que les personnes ne sont pas toutes dans Wikidata, les épisodes de séries ne le sont pas tous également. De plus, si seuls 10% des épisodes ont été alignés, presque 50% des séries d'appartenance de ces épisodes sont alignées. Cela montre que la description faite des séries dans Wikidata s'arrête généralement au niveau de la série, et ne descend que rarement au niveau de l'épisode.

Enfin, l'obtention de l'identifiant Wikidata permet d'obtenir **l'ISAN** des fictions et des épisodes alignés en recherchant pour chaque entité la valeur de la propriété P3212 qui est la propriété accueillant l'ISAN.

[6] Exemple pour la sous-classe Saison: https://query.wikidata.org/#select%20%3Fsaison%20%3FsaisonLabel%0Awhere%7B%0A%3Fsaison%20wdt%3AP31%20wd%3AQ3464665.%0Aservice%20wikibase%3Alabel%7Bbd%3AserviceParam%20wikibase%3Alanguage%20%22fr%2Cen%22%7D%0A%7D