

Présentation du projet DATAIKU

Mobilités et approches comparées entre sciences humaines et sciences dures : la *Revue d'histoire des Sciences/les Annales. Histoire, Sciences sociales* (Persée – Isidore - Wikidata)

Avertissement : Afin de récupérer les données pour chacune des revues comprises dans la collection de la *Revue d'histoire des sciences* et pour celles de la collection des *Annales*, il a fallu répéter un certain nombre de requêtes SPARQL en modifiant à chaque fois uniquement l'identifiant de la revue figurant en première ligne de requête. Afin de ne pas surcharger inutilement notre présentation, nous n'avons – pour chaque exemple – intégré qu'une seule de chacune de ces requêtes.

I – Choix d'une revue et de ses angles d'analyse

Désireux de pouvoir explorer les jeux de données Persée dans toute leur richesse, nous avons choisi la *Revue d'histoire des Sciences* comme source principale. En effet, les nombreuses publications de cette dernière s'étirent sur une vaste étendue chronologique (1947-2006) et leurs thématiques éditoriales interdisciplinaires se prêtent aisément à une approche multifacette et comparée. Il nous a semblé particulièrement opportun de profiter de ce projet pour extraire des données susceptibles de renseigner sur les interactions et mobilités entre sciences humaines et sciences dures, tant au niveau des auteurs qui en sont issus, que dans les productions scientifiques qu'elles génèrent. Nous voulions de cette façon établir une forme de « coupe sociologique » des auteurs en histoire des sciences en déterminant leurs origines socio-professionnelles, leurs niveaux d'études, leurs nationalités, leurs âges

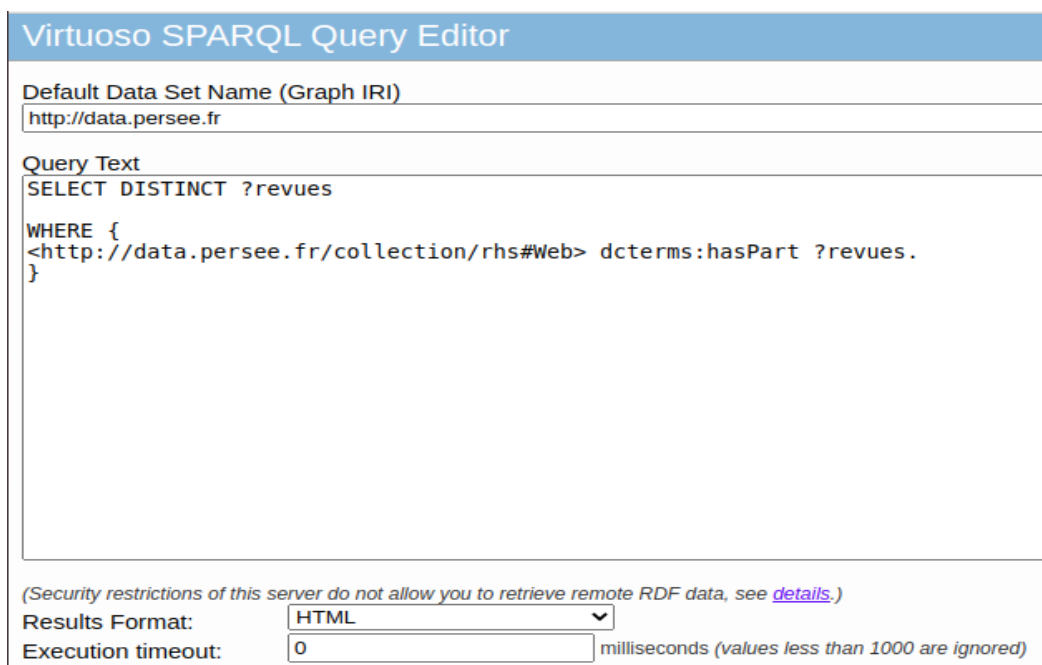
et leurs lieux d'exercices. Afin de rendre ces données plus intelligibles et fécondes, nous les avons comparées « en miroir » avec les jeux de données extraits de la revue des *Annales. Histoire, Sciences sociales*, selon une symétrie d'analyse et de visualisation rigoureuse. Ce choix s'est imposé naturellement car nous recherchions une revue partageant la même ligne historiographique générale que la *RHS* mais appliquée aux sciences humaines et sociales (SHS), et dont la chronologie ainsi que la densité éditoriale seraient relativement proches de celles de notre revue initiale. Cette approche comparée nous a permis de déterminer, d'une part, les points de variation et de rapprochement entre les profils des auteurs en histoire des sciences, et ceux des auteurs en histoire des SHS ; d'autre part, d'établir si les acteurs des sciences dures et ceux des SHS étaient également les auteurs d'une histoire de leurs disciplines respectives.

II – Extraction et croisement des données de la RHS

1. Persée : l'extraction des jeux de données primaires

Dans un premier temps, nous avons extrait nos jeux de données primaires par l'intermédiaire du SPARQL Endpoint de Persée. Au préalable, il a été nécessaire de bien comprendre la structure et la syntaxe de la modélisation en graphe des données. Le téléchargement et l'exploration des jeux de données par fichier « Autorité » et « Collection » ont offert des exemples concrets en révélant les ressources, prédicats et objets indispensables à nos requêtes. Nous avons ainsi compris que la collection de la *Revue d'histoire des Sciences* (bibo : Collection) est reliée à ses deux revues (bibo : Journal) par le prédicat hiérarchiquement descendant 'dcterms : hasPart'. Les revues comprennent elles-mêmes des numéros (bibo : Issue) qui regroupent des articles (bibo : Document). Selon une hiérarchie ascendante, le prédicat les reliant est 'dcterms : isPartOf'. Ce premier état des lieux syntaxique a été indispensable pour comprendre le niveau d'arborescence auquel nous devions effectuer les requêtes susceptibles de

renvoyer les données bibliographiques et numériques inhérentes aux articles, ainsi que celles relatives aux auteurs, en vue de leur croisement ultérieur avec les données d'Isidore et de Wikidata. Afin de suivre la structure de l'arborescence, nous avons effectué une première requête permettant la récupération des deux revues de la *Revue d'histoire des Sciences* (RHS).



Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)
http://data.persee.fr

Query Text
SELECT DISTINCT ?revues
WHERE {
<http://data.persee.fr/collection/rhs#Web> dcterms:hasPart ?revues.
}

(Security restrictions of this server do not allow you to retrieve remote RDF data, see [details](#).)

Results Format: HTML

Execution timeout: 0 milliseconds (values less than 1000 are ignored)

Dans un second temps, nous avons mis au point une requête afin d'obtenir l'intégralité des articles, certaines informations bibliographiques associées ainsi que les auteurs pour la première revue. Une requête parfaitement similaire a été faite ensuite pour récupérer ces données pour la seconde revue. Afin d'atteindre le niveau de l'article, il a d'abord fallu extraire les numéros (?articles dcterms:isPartOf?numeros). Pour chaque article, nous avons souhaité le renvoi des identifiants des auteurs ainsi que leur nom complet via les prédicats 'marcrel:aut' et 'rdfs:label'. Cette donnée nous a semblé plus pertinente et exploitable qu'une information distinguant le nom de famille et le prénom, et nous avons donc délibérément écarté l'utilisation des prédicats associés à cette distinction. Notre requête a également exhumé les titres des articles, leurs identifiants Persée, la responsabilité éditoriale et la date de publication de la version imprimée, l'entité responsable de l'édition numérique et,

optionnellement, la date de publication de l'édition numérique ainsi que l'objet imprimé qui fait l'objet d'une reproduction électronique.

Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)

http://data.persee.fr

Query Text

```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX marcrel: <http://id.loc.gov/vocabulary/relators/>
PREFIX rdam: <http://rdaregistry.info/Elements/m/>
PREFIX rdau: <http://rdaregistry.info/Elements/u/>

SELECT DISTINCT ?numeros ?articles ?auteur ?auteurlabel ?titre ?id ?datepubnum ?print ?datepubimp ?pubnum ?pubimp

WHERE {
  ?numeros dcterms:isPartOf <http://data.persee.fr/collection/rhs_0151-4105#Web>.
  ?articles dcterms:isPartOf ?numeros.
  ?articles marcrel:aut ?auteur.
  ?auteur rdfs:label ?auteurlabel.
  ?articles dcterms:title ?titre.
  ?articles dcterms:identifiant ?id.
  OPTIONAL {?articles rdam:dateOfPublication ?datepubnum.}
  OPTIONAL {?articles rdau:electronicReproductionOf ?print.}
  ?print rdam:dateOfPublication ?datepubimp.
  ?articles dcterms:publisher ?pubnum.
  ?print dcterms:publisher ?pubimp.
}

order by ?auteur
```

(Security restrictions of this server do not allow you to retrieve remote RDF data, see [details](#).)

Results Format:

HTML

Execution timeout:

0

 milliseconds (values less than 1000 are ignored)

Options:

☒ Strict checking of void variables

☐ Log debug info at the end of output (has no effect on some queries and output formats)

☐ Generate SPARQL compilation report (instead of executing the query)

(The result can only be sent back to browser, not saved on the server, see [details](#))

2. Croisement avec les enrichissements d'Isidore et de Wikidata

Nous avons dans un troisième temps élaboré les requêtes destinées à recueillir les données disponibles sur les SPARQL Endpoint de Wikidata, pour enrichir les données sur les auteurs de la *RHS* ; et d'Isidore, pour enrichir les données des articles récupérées avec Persée.

Sur Wikidata, la base de la requête était simple : l'objectif était de récupérer toutes les personnes étant une 'instance of'¹ de la classe 'human'² possédant au moins un identifiant Persée.³ Au sein de la requête, Il ne restait ensuite qu'à utiliser l'opérateur binaire 'OPTIONAL' pour récupérer les informations que nous avons jugé intéressantes, sans pour autant contraindre la requête en éliminant les entrées ne possédant pas ces informations. Nous avons ainsi souhaité récupérer les informations suivantes :

- La date de naissance et la date de décès.⁴ Afin de restreindre la requête et dans le but de l'optimiser, un filtre a été appliqué sur la date de naissance : seules les personnes nées après ou pendant 1850 ont été récupérées. La *RHS* ayant été créée en 1947, il a été jugé que l'année 1850 permettait de récupérer même les auteurs les plus âgés. La date de décès n'a elle pas été filtrée afin de garder les auteurs encore vivants.
- 'Sex or gender,' indiquant ou le genre ou le sexe de la personne.⁵
- 'Country of citizenship,' indiquant le pays où la personne est citoyenne.⁶
- 'Occupation,' indiquant le métier de la personne.⁷
- 'Educated at,' indiquant où la personne a suivi son enseignement.⁸
- 'Academic degree,' indiquant le diplôme universitaire détenu par la personne.⁹
- L'identifiant IdRef, dans l'idée de possiblement enrichir les données plus tard.¹⁰

1 Voir la propriété P31 sur Wikidata, <https://www.wikidata.org/wiki/Property:P31>.

2 Voir la notice Q5 sur Wikidata, <https://www.wikidata.org/wiki/Q5>.

3 Voir la propriété P27732 sur Wikidata, <https://www.wikidata.org/wiki/Property:P2732>.

4 Pour la date de naissance, voir la propriété P569 sur Wikidata, <https://www.wikidata.org/wiki/Property:P569> ; pour la date de décès, voir la propriété P570, <https://www.wikidata.org/wiki/Property:P570>.

5 Voir la propriété P21 sur Wikidata, <https://www.wikidata.org/wiki/Property:P21>.

6 Voir la propriété P27 sur Wikidata, <https://www.wikidata.org/wiki/Property:P27>.

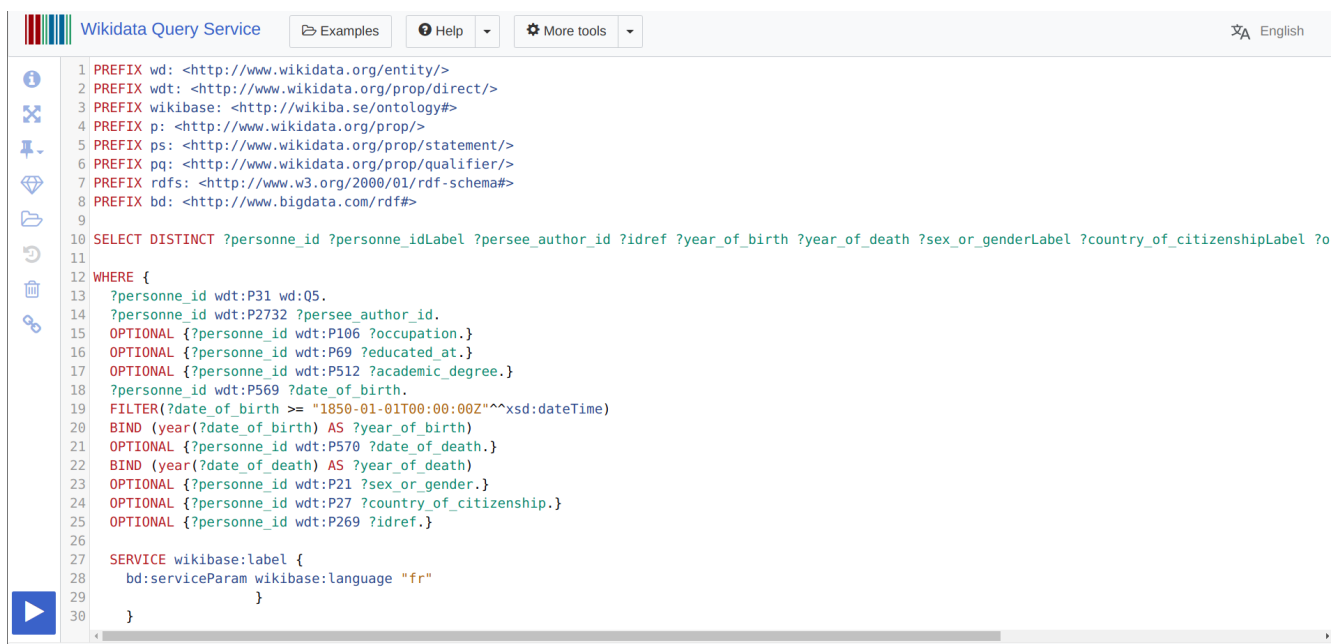
7 Voir la propriété P106 sur Wikidata, <https://www.wikidata.org/wiki/Property:P106>.

8 Voir la propriété P69 sur Wikidata, <https://www.wikidata.org/wiki/Property:P69>.

9 Voir la propriété P512 sur Wikidata, <https://www.wikidata.org/wiki/Property:P512>.

10 Voir la propriété P269 sur Wikidata, <https://www.wikidata.org/wiki/Property:P269>.

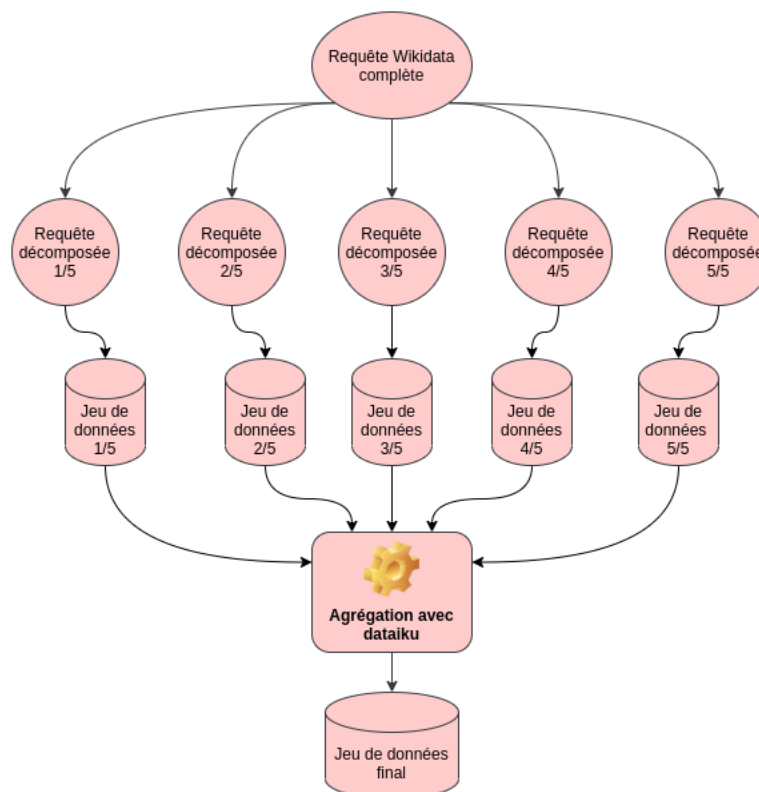
Le SPARQL endpoint de Wikidata offre la possibilité de récupérer les labels de ces propriétés, ainsi que la langue dans laquelle le label doit être récupéré. Pour cela, le français a naturellement été choisi.

The image shows a screenshot of the Wikidata Query Service web interface. At the top, there's a header with the Wikidata logo, the text 'Wikidata Query Service', and several buttons: 'Examples', 'Help', and 'More tools'. On the right, there's a language selector set to 'English'. The main area contains a SPARQL query written in a light blue monospace font. The query starts with several PREFIX declarations for Wikidata entities and properties. It then uses a SELECT DISTINCT statement to retrieve various labels for a specific person, with a WHERE clause containing multiple OPTIONAL and FILTER conditions. At the bottom, there's a SERVICE declaration for Wikibase:label to specify the language as French. A blue play button icon is visible on the left side of the query editor.

```
1 PREFIX wd: <http://www.wikidata.org/entity/>
2 PREFIX wdt: <http://www.wikidata.org/prop/direct/>
3 PREFIX wikibase: <http://wikiba.se/ontology#>
4 PREFIX p: <http://www.wikidata.org/prop/>
5 PREFIX ps: <http://www.wikidata.org/prop/statement/>
6 PREFIX pq: <http://www.wikidata.org/prop/qualifier/>
7 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
8 PREFIX bd: <http://www.bigdata.com/rdf#>
9
10 SELECT DISTINCT ?personne_id ?personne_idLabel ?persee_author_id ?idref ?year_of_birth ?year_of_death ?sex_or_genderLabel ?country_of_citizenshipLabel ?oc
11
12 WHERE {
13   ?personne_id wdt:P31 wd:Q5.
14   ?personne_id wdt:P2732 ?persee_author_id.
15   OPTIONAL {?personne_id wdt:P106 ?occupation.}
16   OPTIONAL {?personne_id wdt:P69 ?educated_at.}
17   OPTIONAL {?personne_id wdt:P512 ?academic_degree.}
18   ?personne_id wdt:P569 ?date_of_birth.
19   FILTER(?date_of_birth >= "1850-01-01T00:00:00Z"^^xsd:dateTime)
20   BIND (year(?date_of_birth) AS ?year_of_birth)
21   OPTIONAL {?personne_id wdt:P570 ?date_of_death.}
22   BIND (year(?date_of_death) AS ?year_of_death)
23   OPTIONAL {?personne_id wdt:P21 ?sex_or_gender.}
24   OPTIONAL {?personne_id wdt:P27 ?country_of_citizenship.}
25   OPTIONAL {?personne_id wdt:P269 ?idref.}
26
27   SERVICE wikibase:label {
28     bd:serviceParam wikibase:language "fr"
29   }
30 }
```

Requête SPARQL wikidata complète, voir annexes.

Malheureusement, une requête demandant autant d'information demandait trop de ressources au client, aboutissant systématiquement sur une erreur 'Query timeout limit reached' ou à une erreur du navigateur internet. Afin de contourner ce problème, il a été décidé de décomposer cette requête en cinq petites requêtes, puis de les agréger grâce à Dataiku. Ces cinq requêtes ont été jointes en annexes.



Pour le SPARQL endpoint d'Isidore, la requête a été construite autour du modèle de graphe RDF spécifique à la plate-forme.¹¹ Sur Isidore, les articles sont rassemblés dans des collections grâce au prédicat 'ore:isAggregatedBy', la collection n'étant autre que la *RHS* dans notre cas. Ainsi, il suffisait de récupérer chaque article agrégé par la collection « Revue d'histoire des sciences ».¹² A cela est venu s'ajouter les informations suivantes :

- Son identifiant, 'dcterms:identifiant,' permettant à terme de croiser le jeu de données avec les données de Persée.
- Son auteur.ice, 'dcterms:creator.'
- Son titre, 'dcterms:title.'
- Sa date, 'dc:date.'

11 Voir le modèle de graphe d'Isidore orienté sur le document: <https://isidore.science/img/isidore-document-model.png>.

12 Voir l'identifiant de la collection en question : <http://isidore.science/collection/10670/2.eqird7>.

- Son type, 'dc:type,' qui correspond au type de l'oeuvre écrite, un article ou un compte-rendu, par exemple.
- Son 'coverage,' 'dc:coverage,' qui correspond à l'intervalle de pages concerné par l'oeuvre écrite.
- La collection qui agrège l'oeuvre écrite récupérée, 'ore:agregates.'
- Son sujet, 'dc:subject,' accompagné de l'opérateur binaire OPTIONAL, pour ne pas supprimer les entrées n'ayant pas cette information.
- Enfin, une dernière information discriminante a été récupérée : le 'topic,' 'sioc:topic.' Pour ne pas supprimer de résultat, et à cause de la nature de la requête, où le label du topic ne peut être récupéré qu'en deux temps, deux requêtes ont été nécessaires. Une pour les articles avec topic, et une pour les articles sans. Pour cela, un filtre a été utilisé.


```

PREFIX sioc: <http://rdfs.org/sioc/ns#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX ore: <http://www.openarchives.org/ore/terms/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

SELECT DISTINCT ?id (SAMPLE(?author) AS ?author) ?title ?date ?type ?topic_label ?subject ?coverage ?is_aggregated_by

WHERE {
?documents ore:isAggregatedBy <http://isidore.science/collection/10670/2.eqird7>.
?documents dcterms:identifier ?id.
?documents dcterms:creator ?name.
?name foaf:name ?author.
?documents dcterms:title ?title.
?documents dc:date ?date.
?documents dc:type ?type.
?documents sioc:topic ?topic.
OPTIONAL {?topic skos:prefLabel ?topic_label.}
FILTER(lang(?topic_label)="fr")
OPTIONAL {?documents dc:subject ?subject}
?documents dc:coverage ?coverage.
?is_aggregated_by ore:aggregates ?documents.
}

```

Requête SPARQL Isidore pour les articles avec topic

```

PREFIX sioc: <http://rdfs.org/sioc/ns#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX ore: <http://www.openarchives.org/ore/terms/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

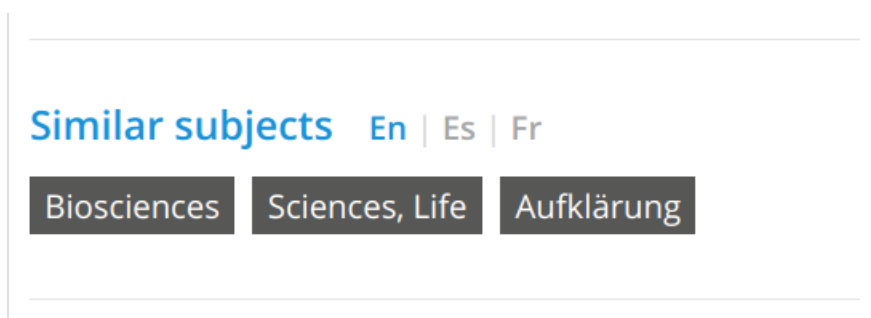
SELECT DISTINCT ?id (SAMPLE(?author) AS ?author) ?title ?date ?type ?subject ?is_aggregated_by ?coverage

WHERE {
?documents ore:isAggregatedBy <http://isidore.science/collection/10670/2.eqird7>.
?documents dcterms:identifier ?id.
?documents dcterms:creator ?name.
?name foaf:name ?author.
?documents dcterms:title ?title.
?documents dc:date ?date.
?documents dc:type ?type.
FILTER NOT EXISTS {?documents sioc:topic ?topic.}
OPTIONAL {?documents dc:subject ?subject}
?is_aggregated_by ore:aggregates ?documents.
?documents dc:coverage ?coverage.
}

```

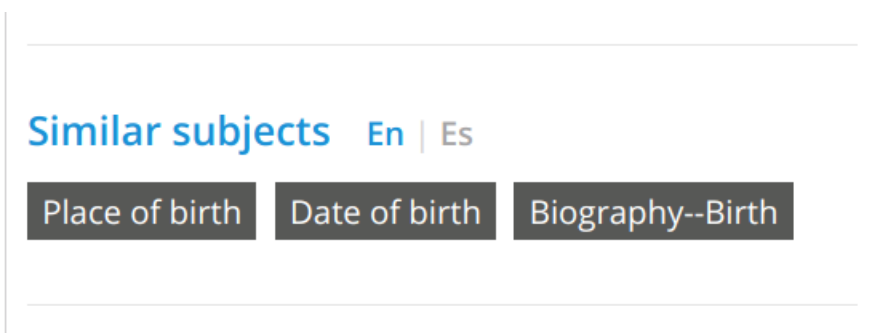
Requête SPARQL Isidore pour les articles sans topic

L'information des sujets ('subject') étant relativement peu présente et donc difficilement exploitable, nous aurions souhaité pouvoir récupérer les mots-clés indiqués dans la rubrique 'Similar subjects' des différentes notices de chaque article. Par exemple, pour l'article « Roselyne Rey (1951-1995) », écrit par D. Gourevitch, A. F. Largeault et A. M. Chouillet pour la *Revue d'histoire des sciences* en 1995¹³, les mots-clés suivants sont indiqués :



Cependant, et d'après nos recherches, le modèle de graphe d'Isidore ne permet pas de récupérer cette information. Le N3/Turtle de cette notice ne contient en effet pas ces informations, et aucun prédicat indiqué sur le diagramme présentant le modèle de graphe n'indique où récupérer cette donnée.

En outre, nous avons relativisé cette absence d'information en réalisant que pour certaines notices, les mots-clés n'étaient pas pertinents et n'auraient pas pu faire l'objet d'une visualisation de données cohérente. Par exemple, pour l'article « Arago et la naissance de la polarimétrie », écrit par J. Rosmorduc pour la *Revue d'histoire des sciences* en 1988,¹⁴ les mots-clés sont les suivants :



¹³ Voir <https://isidore.science/document/10670/1.h19xpb>.

¹⁴ Voir <https://isidore.science/document/10.3406/rhs.1988.4087>.

III - Nettoyage et agrégation des données

Une fois les différents jeux de données obtenus, nous avons pu les nettoyer et les agréger grâce à Dataiku. Trois projets ont été créés :

- RHS Authors – Persée. Ce projet avait pour but de créer un jeu de données ne contenant que les noms et les identifiants des auteurs de la *RHS*, à partir du jeu de données récupéré via Persée.
- RHS Authors – Persée & Wikidata. Ce projet agrège les différentes requêtes Wikidata, tout en veillant à effectuer un dédoublement à chaque étape. Les données récupérées ont ensuite été croisées, grâce à une jointure faite avec l'identifiant Persée, avec le jeu de données contenant les identifiants Persée et les noms des auteurs de la *RHS*. Une fois agrégés, les colonnes inutiles ont été enlevées et chaque cellule vide a été remplie par la valeur « no value », indiquant une absence d'information. Pour les dates de décès absentes, quatre points d'interrogation, « ???? », sont venus remplir les cellules vides. Du jeu de données rassemblant toutes les informations récupérées sur Wikidata pour chaque auteur, quatre jeux de données finaux ont été créés dans le but de réaliser les visualisations de données. Un premier étant simplement le même, au cas où vous voudriez continuer à sélectionner des colonnes ; un deuxième rassemblant les informations sur le pays de citoyenneté ainsi que le niveau de diplôme obtenu ; un troisième rassemblant les dates de naissance et de décès ainsi que le genre ou le sexe. Ce jeu de données est important car ne contenant aucun doublon induit par la répétition d'une information, par exemple le métier, ou « occupation ». Il permet de mesurer justement la répartition des sexes/genres au sein de la revue. Un quatrième se concentre sur la colonne des occupations.
- RHS Articles – Persée & Isidore. Ce projet agrège les informations récupérées via Persée et Isidore sur les articles. Dans un premier temps, les deux jeux de données récupérés via Persée

pour les deux revues différentes de la *RHS* ont été agrégés, puis le résultat a été préparé. Parallèlement, les deux jeux de données obtenus via Isidore pour les articles avec et sans topic ont été préparés, puis agrégés. Deuxièmement, les deux résultats ont été rassemblés grâce à une jointure faite sur l'identifiant Persée de l'article. Troisièmement, les informations de genre ainsi que la date de naissance et de décès de chaque auteur.ice ont été ajoutées pour chaque article, permettant de possibles visualisation de données.

Il est important de préciser que, pour le projet « RHS Authors – Persée & Wikidata », nous passons de 749 rangées, et donc 749 auteurs, à 333 rangées en bout de chaîne, pour le jeu de données ne contenant aucun doublon possédant les informations de sexe/genre. Cette perte de données nous échappe malheureusement. Nous avons cependant trouvé plusieurs causes à celle-ci.

Premièrement, nous avons remarqué que certaines notices d'auteur.ice disponibles sur Persée possédaient des doublons, avec un identifiant et des informations différentes. Cela empêche de faire le lien entre Wikidata et Persée lorsque l'identifiant Persée récupéré sur Wikidata n'est pas le même que celui récupéré via la requête initiale des articles via Persée. Par exemple, pour l'auteur Raymond Furon, on retrouve grâce à notre requête SPARQL sur Persée l'information suivante :

auteur_persee_id	auteurlabel
string Integer	string Text
7583	Raymond Furon

Tandis qu’avec notre requête SPARQL sur Wikidata, l’identifiant Persée diffère de la sorte :

personne_id	personne_idLabel	persee_author_id
string URL	string Text	string Integer
http://www.wikidata.org/entity/Q55772512	Raymond Furon	386941

Les deux notices existent de plus bel et bien sur Persée, leur RDF également.¹⁵ De plus, le fait qu’il n’existe qu’une seule notice consacrée à Raymond Furon sur Wikidata montre qu’il ne s’agit très probablement pas d’un problème d’homonymie.

Deuxièmement, nous avons remarqué que certaines notices wikidata des auteurs de la *RHS* n’indiquent pas d’identifiant Persée, la jointure réalisée grâce à celui-ci n’est donc pas possible. Nous avons souhaité maintenir la jointure avec l’identifiant, jugeant qu’une jointure sur le nom était trop hasardeuse à cause des possibles variations (prénom abrégé, prénom et nom inversés, etc.).¹⁶

IV – Extraction, croisement et traitement des jeux de données secondaires

1. Extraction des jeux de données secondaires

Après l’extraction, l’agrégation et le traitement des données inhérentes aux auteurs et aux articles de la *RHS*, nous avons réitéré le processus pour les *Annales. Histoire, Sciences sociales* (désormais *AHSS*). Toutefois, nous nous sommes limités pour cette revue aux données nécessaires à la comparaison « socio-professionnelle » et avons donc écarté les informations relatives au type de contenu documentaire. Les requêtes effectuées sur SPARQL Endpoint de Persée ont suivi les mêmes étapes successives que celles réalisées pour les données de la *RHS*. Ainsi, nous avons d’abord récupéré les revues contenues dans la collection des *AHSS* par la requête suivante :

¹⁵ Voir <https://www.persee.fr/authority/7583> et <https://www.persee.fr/authority/386941>.

¹⁶ Par exemple, pour Fabien Chareix, un des auteurs de la *Revue d’histoire des sciences*, on retrouve sa notice Persée ici : <https://www.persee.fr/authority/10464>, et donc un identifiant Persée, mais sa notice wikidata <https://www.wikidata.org/wiki/Q104628669> ne contient pas cette information.

Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)

<http://data.persee.fr>

Query Text

```
SELECT DISTINCT ?revues
```

```
WHERE {  
<http://data.persee.fr/collection/ahess#Web> dcterms:hasPart ?revues.  
}
```

Afin de récupérer les auteurs des *AHSS* et leurs identifiants, nous avons répété la requête utilisée pour ceux de la *RHS*, pour chacune des cinq revues sur les six que contient la collection des *Annales*. La première revue, qui comprend les numéros publiés entre 1929 et 1942 n'a pas été prise en compte car antérieure à la création de la *RHS*, ce qui aurait éventuellement biaisé les données de comparaison.

Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)

http://data.persee.fr

Query Text

```
SELECT DISTINCT ?numeros ?articles ?titre ?auteur ?auteurlabel

WHERE {
  ?numeros dcterms:isPartOf <http://data.persee.fr/collection/ahess_1243-2563#Web>.
  ?articles dcterms:isPartOf ?numeros.
  ?articles dcterms:title ?titre.
  ?articles marcrel:aut ?auteur.
  ?auteur rdfs:label ?auteurlabel.
}

order by ?auteur
```

(Security restrictions of this server do not allow you to retrieve remote RDF data, see [details](#).)

Results Format:

HTML

Execution timeout:

0

milliseconds (values less than 1000 are ignored)

Options:

☒ Strict checking of void variables

Un autre jeu de données secondaires, dédié à la connaissance de l'ensemble des revues pour lesquelles les auteurs de la *RHS* sont également contributeurs, a également été récupéré sur Persée grâce à la requête suivante :

Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)

http://data.persee.fr

Query Text

```
SELECT DISTINCT ?numeros ?articles ?id ?titre ?auteur ?auteurlabel ?articlerevues ?num ?revue ?label

WHERE {
  ?numeros dct:isPartOf <http://data.persee.fr/collection/rhs_0151-4105#Web>.
  ?articles dct:isPartOf ?numeros.
  ?articles dct:title ?titre.
  ?articles dct:identifier ?id.
  ?articles marcel:aut ?auteur.
  ?auteur rdfs:label ?auteurlabel.
  ?articlerevues marcel:aut ?auteur.
  ?articlerevues dct:isPartOf ?num.
  ?num dct:isPartOf ?revue.
  ?revue rdfs:label ?label.
}
```

(Security restrictions of this server do not allow you to retrieve remote RDF data, see [details](#).)

Results Format:

HTML

Execution timeout:

0

milliseconds (values less than 1000 are ignored)

Options:

☒ Strict checking of void variables

☐ Log debug info at the end of output (has no effect on some queries and output formats)

☐ Generate SPARQL compilation report (instead of executing the query)

(The result can only be sent back to browser, not saved on the server, see [details](#))

Run Query

Reset

Sachant que nous souhaitons récupérer tous les articles et les noms de revues dont ils sont extraits pour chaque auteur de la *RHS*, nous avons extrapolé la requête utilisée pour le jeu de données primaires. Ainsi, par un jeu de requête en cascade, nous avons utilisé le prédicat ‘marcel:aut’ pour extraire tous les articles ayant pour auteur l’un des auteurs de la *RHS* récupérés grâce à la ligne précédente. Puis nous avons remonté l’arborescence jusqu’au niveau de la revue, afin de récupérer le label de chacune d’elle. De cette façon, nous avons obtenu tous les articles recensés dans Persée pour chaque auteur de la *RHS* ainsi que le nom des revues associées.

16/20

Afin de mettre au jour les dynamiques historiographiques qui sous-tendent les publications de la *RHS*, nous avons vérifié l'éventuelle prédominance des sciences biologiques et médicales en identifiant la part des auteurs de la RHS qui ont également contribué à la *Revue d'Histoire de la Pharmacie* ainsi que leurs professions. Nous avons ainsi extrait un troisième jeu de données supplémentaire par l'intermédiaire d'une requête SPARQL sur Persée, réalisée pour chacune des deux revues de la collection *RHS* :

The screenshot shows the Virtuoso SPARQL Query Editor interface. At the top, there's a blue header with the text "Virtuoso SPARQL Query Editor". Below this, there's a section for "Default Data Set Name (Graph IRI)" with a text input field containing "http://data.persee.fr". Underneath is a section for "Query Text" with a large text area containing a SPARQL query. The query starts with three prefixes: "PREFIX dcterms: <http://purl.org/dc/terms/>", "PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>", and "PREFIX marcrel: <http://id.loc.gov/vocabulary/relators/>". The main query is a SELECT statement with DISTINCT, selecting variables ?numeros, ?articles, ?id, ?titre, ?auteur, ?auteurlabel, and ?articlerevues. The WHERE clause includes several conditions: ?numeros dcterms:isPartOf <http://data.persee.fr/collection/rhs_0151-4105#Web>, ?articles dcterms:isPartOf ?numeros, ?articles dcterms:title ?titre, ?articles dcterms:identifiant ?id, ?articles marcrel:aut ?auteur, ?auteur rdfs:label ?auteurlabel, and ?articlerevues marcrel:aut ?auteur. There is also a FILTER regex(?articlerevues, 'pharm'). The query ends with "order by ?auteur". At the bottom, there's a note about security restrictions and a "Results Format" dropdown menu set to "HTML".

```

Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)
http://data.persee.fr

Query Text
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX marcrel: <http://id.loc.gov/vocabulary/relators/>

SELECT DISTINCT ?numeros ?articles ?id ?titre ?auteur ?auteurlabel ?articlerevues
WHERE {
  ?numeros dcterms:isPartOf <http://data.persee.fr/collection/rhs_0151-4105#Web>.
  ?articles dcterms:isPartOf ?numeros.
  ?articles dcterms:title ?titre.
  ?articles dcterms:identifiant ?id.
  ?articles marcrel:aut ?auteur.
  ?auteur rdfs:label ?auteurlabel.
  ?articlerevues marcrel:aut ?auteur.
  FILTER regex(?articlerevues, 'pharm')
}

order by ?auteur

(Security restrictions of this server do not allow you to retrieve remote RDF data, see details.)
Results Format: HTML

```

La structure de cette requête se décompose de cette façon : pour chaque numéro contenu dans la revue de la *RHS* sélectionnée au moyen de l'identifiant récupéré par la requête initiale (cf. extraction des jeux de données primaires), nous souhaitons le renvoi de chaque article, son titre, son identifiant, le nom de l'auteur et l'identifiant numérique. Puis nous requêtons sur l'ensemble des articles recensés

dans la base de Persée dont les auteurs sont aussi ceux de la RHS. Nous filtrons ensuite les résultats en utilisant une expression régulière (Regex), après avoir examiné la structure centrale - 'pharm' - de la ressource identifiante de la *Revue d'histoire de la Pharmacie*. Les résultats ont été probants et n'ont pas généré de « bruit » excessif.

2. Nettoyage et agrégation des jeux de données secondaires

Pour le premier jeu de données secondaires, la même méthodologie utilisée pour le traitement des jeux de données primaires a été appliquée pour agréger le jeu de données Wikidata et le jeu de données des AHSS récupéré via Persée (Cf projet Dataiku « Annales – Persée & Wikidata »).

Concernant le second jeu récupérant l'ensemble des articles recensés dans Persée pour chaque auteur de la RHS ainsi que les labels des revues, il n'a nécessité que peu de traitement. Nous avons réalisé une suppression des doublons par la fonctionnalité Dataiku 'Distinct' et n'avons conservé que les colonnes inhérentes aux ressources identifiantes des articles et aux labels des revues dont ces derniers sont extraits. En effet, ces deux types de données étaient suffisants pour offrir une visualisation proportionnée de la part des différentes revues auxquelles contribuent les auteurs de la RHS. Le total de nombre d'articles pour chaque revue en permet l'analyse volumétrique.

Enfin, le troisième jeu de données obtenu sur la *Revue d'Histoire de la Pharmacie* a été croisé avec les données Wikidata pour obtenir les professions des auteurs de la RHS qui ont également publié dans la RHP. Il a bénéficié des traitements réalisés sur les jeux de données primaires puis a fait l'objet d'une extraction à part avec la fonctionnalité Dataiku « filter ».

V – Datavisualisations : choix narratifs et graphiques

Suite à l'agrégation et au traitement de nos données primaires et secondaires, nous avons procédé à leur exploitation sous forme de visualisation de données en utilisant le logiciel Tableau Public. Ce dernier permet l'élaboration d'« histoire », que nous avons décomposée en deux grands chapitres distincts : « Données pour une sociologie des auteurs de la *Revue d'histoire des Sciences* » et « Données pour une historiographie de la *Revue d'histoire des Sciences* ». Le modèle du diagramme circulaire a été privilégié lorsque les catégories de résultats étaient peu nombreuses et pouvaient donc permettre une visualisation lisible et parfaitement explicite. Il en a été ainsi pour la visualisation de la proportion des auteurs par genre (*RHS* et *AHSS*) et pour la part des auteurs de la *RHS* et de leurs professions dans la *Revue d'histoire de la Pharmacie*. Le modèle du diagramme en barres verticales nous a semblé quant à lui approprié pour rendre compte avec une forte distinction chromatique et visuelle de la répartition des auteurs par âge, dans une perspective comparée entre ceux de la *RHS* et ceux des *AHSS*. Cette comparaison en intègre une autre, temporelle cette fois-ci, sur la base de deux années retenues (1951 et 2005) pour analyser l'évolution entre les débuts de la revue concernée et sa vie éditoriale récente. Ce modèle a également été utile pour la répartition par type des publications de la *RHS*. Le diagramme en barres horizontales est prédominant dans notre narration en raison de la forte richesse des labels associés aux résultats de certains champs. Nous l'avons utilisé pour visualiser les professions les plus représentées parmi les auteurs de la *RHS* et des *AHSS*, leurs titres et diplômes, et leurs nationalités. Le diagramme « en bulles » a lui été mis à profit lorsque nous souhaitions mettre en relief les données en fonction de leurs proportions, permettant une mise en évidence des labels prédominants. Il en a été ainsi pour la présentation des institutions et organismes d'exercice les plus représentés parmi les auteurs de la *RHS*, dans le cadre d'une comparaison avec ceux des *AHSS*. Enfin, nous avons choisi un diagramme par « briques » afin d'organiser la visualisation proportionnée sous

forme d'inventaire gradué de la part des autres revues Persée ayant les auteurs de la RHS pour contributeurs.

La narration offerte sur Tableau Public au gré des datavisualisations a mis au jour les dynamiques et zones de contraste recherchées entre les sciences humaines et les sciences dures, en déconstruisant partiellement les présupposés de genre et de professions inhérents à l'histoire respective des deux champs. Une très large prédominance masculine s'impose ainsi autant en histoire des sciences humaines et sociales qu'en histoire des sciences appliquées. Nous avons pu constater que les acteurs des sciences dures ne sont que minoritairement engagés dans le processus historiographique de leurs disciplines d'activité, permettant *de facto* l'hégémonie des historiens, des philosophes et des écrivains parmi les auteurs de la RHS. Rien d'étonnant donc, à ce que les données socio-professionnelles relatives aux auteurs soient relativement semblables entre la RHS et les AHSS. Il semblerait ainsi que l'histoire des sciences demeure, non sans ironie, l'apanage des sciences humaines, comme une sorte de prolongement naturel et complémentaire.