



HAL
open science

L'agrégation de données

Morgane Pica

► **To cite this version:**

Morgane Pica. L'agrégation de données. Doctorat. Ateliers de l'ARHN, Maison des Sciences de l'Homme, Lyon, France. 2023, pp.37. halshs-04802192

HAL Id: halshs-04802192

<https://shs.hal.science/halshs-04802192v1>

Submitted on 25 Nov 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Cycle Data de l'ARHN
Atelier n°3

L'agrégation de données

Morgane Pica

Ingénieure d'étude
Axe de Recherche en Histoire Numérique
ENS de Lyon
LARHRA UMR 5190

morgane.pica@ens-lyon.fr



LABORATOIRE DE RECHERCHE
HISTORIQUE RHÔNE-ALPES

Plan de l'exposé initial

- 1) Introduction
- 2) Matériel et connaissances prérequis.
- 3) Méthodologie et réflexion sur nos données
- 4) Discussion/questions.

1 – Introduction

Qu'est-ce-qu'un mashup ?

Agrégation de données ?

Cf article et cours de
Gautier Poupeau sur le
blog « Les Petites Cases » :



Tristan Eaton, 4-6 rue du Chevaleret, Paris 13
CC-BY <https://www.flickr.com/photos/lespetitescases/29003193065/>

Un mashup est une œuvre/création originale mise au point à partir de l'assemblage/mise en relation d'œuvres/créations existantes.

Le mot mashup est le plus souvent utilisé pour la musique mais on le retrouve dans les autres arts ainsi que dans le monde numérique lorsqu'on parle de mashup de données.

[https://www.lespetitescases.net/
realiser-mashup-donnees-Dataiku-DSS-Palladio](https://www.lespetitescases.net/realiser-mashup-donnees-Dataiku-DSS-Palladio)

1 – Introduction

Agrégation de données ?

- Aussi appelé *mashup* (en anglais, « mélange »).
- La mise en commun de données issues de sources différentes.
- La création d'une base de données à base de plusieurs autres.
- « Assemblage », « mise en relation » (Cf G.Poupeau).

1 – Introduction

Agrégation de données ?

- La plupart du temps, les données sont d'une manière ou d'une autre complémentaires : leur combinaison peut permettre de compléter l'une, ou d'obtenir de nouvelles informations.
- Typiquement, en SHS, les bases de données peuvent faire référence aux mêmes personnes/événements, mais selon des problématiques différentes : les données enregistrées sont différentes.

1 – Introduction

Agrégation de données ?

- L'agrégation de données sous-entend leur *comparaison* initiale, pour un diagnostic et une prévision du travail.
- Il faudra également *aligner* les modèles entre eux :
 - ↳ quelles sont les données identiques, sur lesquelles se baser (nom d'une personne, date...)?
 - ↳ les données d'un modèle peuvent-elles être transcrites dans l'autre ?...

1 – Introduction

Agrégation de données ?

→ Des décisions s'imposeront :

- ↳ quelle base de données accueillera les autres en sus ? (donc lesquelles transformer?)
- ↳ en informatique on parle souvent de *left-join* ou *right-join*.
- ↳ faut-il complètement repenser le modèle ? (données FAIR)

2 – Prérequis

Matériel

- Un ordinateur,
- Un navigateur et accès internet,
- Un outil de transformation/extraction de données
 - ↳ Dataiku DSS
 - ↳ Langage de programmation
 - ↳ Data Wrangler
- Des données.

2 – Prérequis

Connaissances

- Mon introduction.
- La connaissance de votre champ disciplinaire.
- La connaissance au moins du domaine que concernent les données visées.
- Un but à atteindre (vous vous perdrez moins dans du « au cas où »).

3 – Méthodologie

Questions à se poser :

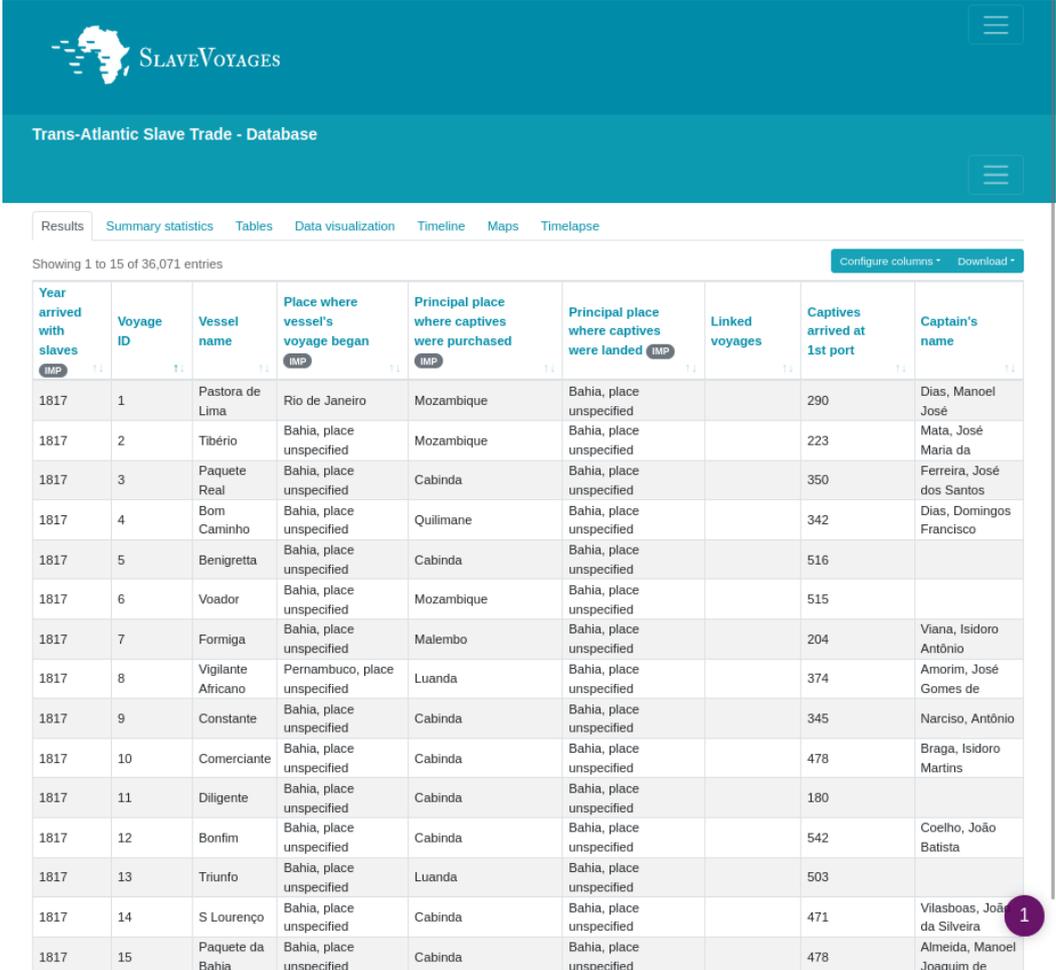
- **Pourquoi** cette agrégation ? (Buts à atteindre.)
- **D'où** partons-nous ? (État initial.)
- Quel état final puis-je **viser** étant donné l'état actuel ?
- Combien de **temps** puis-je y consacrer ?
- Jusqu'à quel point puis-je **automatiser** ?

3 – Méthodologie

De quoi part-on ?

→ Slave Voyages :

- ↳ base sur le commerce triangulaire ;
- ↳ format CSV ;
- ↳ données centrées sur le trajet, aucune autre identification.



The screenshot shows the 'Slave Voyages' database interface. At the top, there is a logo for 'SLAVEVOYAGES' and the title 'Trans-Atlantic Slave Trade - Database'. Below the header, there are navigation tabs: 'Results', 'Summary statistics', 'Tables', 'Data visualization', 'Timeline', 'Maps', and 'Timelapse'. A 'Showing 1 to 15 of 36,071 entries' indicator is present, along with 'Configure columns' and 'Download' options. The main content is a table with the following columns: 'Year arrived with slaves', 'Voyage ID', 'Vessel name', 'Place where vessel's voyage began', 'Principal place where captives were purchased', 'Principal place where captives were landed', 'Linked voyages', 'Captives arrived at 1st port', and 'Captain's name'. The table contains 15 rows of data, each representing a voyage from 1817.

Year arrived with slaves	Voyage ID	Vessel name	Place where vessel's voyage began	Principal place where captives were purchased	Principal place where captives were landed	Linked voyages	Captives arrived at 1st port	Captain's name
1817	1	Pastora de Lima	Rio de Janeiro	Mozambique	Bahia, place unspecified		290	Dias, Manoel José
1817	2	Tibério	Bahia, place unspecified	Mozambique	Bahia, place unspecified		223	Mata, José Maria da
1817	3	Paquete Real	Bahia, place unspecified	Cabinda	Bahia, place unspecified		350	Ferreira, José dos Santos
1817	4	Bom Caminho	Bahia, place unspecified	Quilimane	Bahia, place unspecified		342	Dias, Domingos Francisco
1817	5	Benigretta	Bahia, place unspecified	Cabinda	Bahia, place unspecified		516	
1817	6	Voador	Bahia, place unspecified	Mozambique	Bahia, place unspecified		515	
1817	7	Formiga	Bahia, place unspecified	Malembo	Bahia, place unspecified		204	Viana, Isidoro Antônio
1817	8	Vigilante Africano	Pernambuco, place unspecified	Luanda	Bahia, place unspecified		374	Amorim, José Gomes de
1817	9	Constante	Bahia, place unspecified	Cabinda	Bahia, place unspecified		345	Narciso, Antônio
1817	10	Comerciante	Bahia, place unspecified	Cabinda	Bahia, place unspecified		478	Braga, Isidoro Martins
1817	11	Diligente	Bahia, place unspecified	Cabinda	Bahia, place unspecified		180	
1817	12	Bonfim	Bahia, place unspecified	Cabinda	Bahia, place unspecified		542	Coelho, João Batista
1817	13	Triunfo	Bahia, place unspecified	Luanda	Bahia, place unspecified		503	
1817	14	S Lourenço	Bahia, place unspecified	Cabinda	Bahia, place unspecified		471	Vilasboas, João da Silveira
1817	15	Paquete da Bahia	Bahia, place unspecified	Cabinda	Bahia, place unspecified		478	Almeida, Manoel Joaquim de

3 – Méthodologie

De quoi part-on ?

→ Maritime History :

- ↳ base sur les registres de la Compagnie des Indes de l'Est ;
- ↳ format RDF ;
- ↳ données atomisées, l'analyse de la base peut donc être faite de plusieurs manières.

The screenshot shows the 'Maritime History' SPARQL query results page. At the top right, there are links for 'SPARQL' and 'Search'. Below the title, there is a search bar with the text 'Q Search' and a result count of '9994 Entities'. A 'Class Filter' section is visible, listing various classes with their counts and a checkmark next to 'All classes (23)'. The main results table lists several 'Person' entities, including Jan Ravensbergh, Andries Liens, Abraham Staelboom, Abraham Franke, Jan Pietersz, Jurgen Preen, Andries Jansz., Jan Loker, Isaak Storm, and Jan Albertsz. The 'Isaak Storm' entry is highlighted. At the bottom right, there is a pagination indicator: 'Page 1 of 1000: < < > >'.

3 – Méthodologie

Ce qu'on cherche à faire.

- Deux bases sur le commerce maritime à l'époque moderne.
- Comparer les chiffres disponibles.
- Voir si des capitaines ou navires ont pu changer de route au cours de leurs carrières.

3 – Méthodologie

Problèmes de format.

→ Aligner sur le RDF ?

- ↳ Les avantages du RDF : atomisation, identification systématique, liaison avec le Web sémantique.
- ↳ Atomisation possible.
- ↳ Identification compliquée, voire impossible. Solution : ce que Maritime History propose (dans le doute, une entité par occurrence).
- ↳ Liaison avec le Web sémantique : sur les noms de lieux, possiblement les personnes et vaisseaux.

3 – Méthodologie

Problèmes de format.

→ Aligner sur le CSV ?

- ↳ Les avantages du CSV : plus rapide à faire.
- ↳ Les inconvénients du CSV :
 - ↳ Ici, pas de standard utilisé, donc pas d'interopérabilité.
 - ↳ Identification facultative.

3 – Méthodologie

Problèmes de format.

→ Dans l'idéal :

- ↳ Faire un RDF avec le tableau CSV.
- ↳ Utiliser le même standard (CIDOC-CRM et extensions) que Maritime History, pour des données FAIR.
- ↳ Marquer les possibles correspondances repérées par des liens non définitifs.

3 – Méthodologie

Problèmes d'alignement

- Un standard international (CIDOC-CRM) recherchant l'objectivité **vs** un format « maison » totalement subjectif.
- Peut-on aligner le format « maison » sur le standard international ?
- Peut-on utiliser le standard international sur un tableau ? (Oui!)

3 – Méthodologie

Problèmes d'alignement

→ Dans l'idéal :

- ↳ Faire un RDF avec le tableau CSV.
- ↳ Utiliser le même standard (CIDOC-CRM et extensions) que Maritime History, pour des données FAIR.
- ↳ Marquer les possibles correspondances repérées par des liens non définitifs.

→ Un standard international (CIDOC-CRM) recherchant l'objectivité vs un format « maison » totalement subjectif.

3 – Méthodologie

Problèmes d'alignement

→ Même information ?

↳ « Captain's name »

↳ « Vessel name »

↳ « Vessel owner »

3 – Méthodologie

Problèmes d'alignement

→ Même information ?

↳ « Captain's name »

↳ « Vessel name »

↳ « Vessel owner »

→ Sur Maritime History, on a des personnes et des objets, liées à une ou plusieurs appellations et à des fonctions, ce qui rend les données moins subjectives : on encode d'une manière qui ne dépend pas de la problématique.

3 – Méthodologie

Problèmes d'alignement

→ Même information ?

↳ « Flag of vessel 1 »

↳ « Flag of vessel 2 »

→ Y a-t-il une hiérarchie entre ces deux champs ?

3 – Méthodologie

Problèmes d'alignement

→ Même information ?

- ↳ « Date vessel departed with captives »
- ↳ « Date vessel departed for homeport »
- ↳ « Date vessel arrived with captives »
- ↳ « Date captive embarkation began »
- ↳ « Date vessel's voyage began »
- ↳ « Year constructed »

3 – Méthodologie

Problèmes d'alignement

→ Même information ?

- ↳ « Date vessel departed with captives »
- ↳ « Date vessel departed for homeport »
- ↳ « Date vessel arrived with captives »
- ↳ « Date captive embarkation began »
- ↳ « Date vessel's voyage began »
- ↳ « Year constructed »

→ Ces champs sont ambigus...

3 – Méthodologie

Pouvoir distinguer la provenance.

→ Si CSV : colonne de provenance.

→ Si RDF :

↳ garder l'identifiant d'origine ?

4 – Sur nos données

Les capitaines

- Testons en « exact match ».
- Comparons les prénoms et noms de familles séparément, en ignorant ce qu'il y a au milieu.
- Ajoutons les dates.

3 – Méthodologie

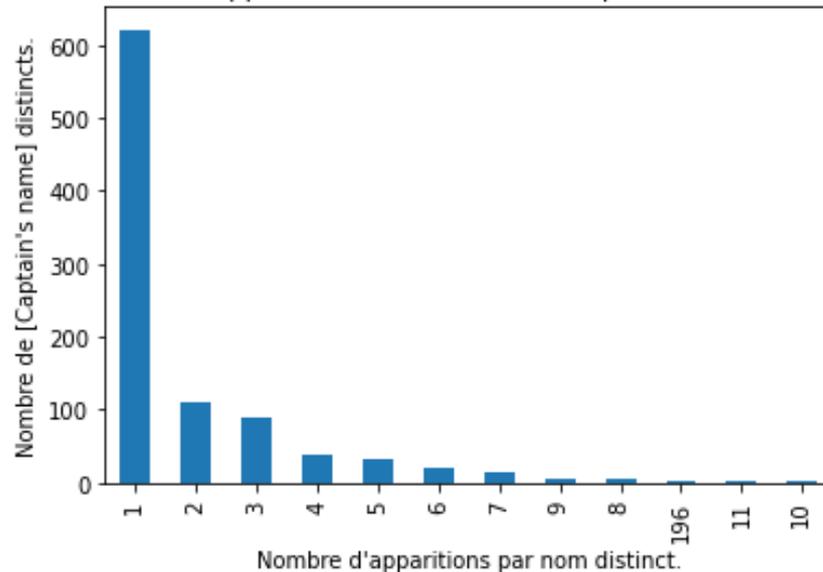
Les capitaines : exact match.

Il y a 1911 lignes au tableau `Slave Voyages`, avec 932 `"Captain's name"` différents.

Comment lire ce diagramme. De gauche à droite :

- 621 `"Captain's name"` appear 1 time.
- 109 `"Captain's name"` appear 2 times.
- 88 `"Captain's name"` appear 3 times.
- 38 `"Captain's name"` appear 4 times.
- 32 `"Captain's name"` appear 5 times.
- 20 `"Captain's name"` appear 6 times.
- 13 `"Captain's name"` appear 7 times.
- 4 `"Captain's name"` appear 9 times.
- 4 `"Captain's name"` appear 8 times.
- 1 `"Captain's name"` appears 196 times.
- 1 `"Captain's name"` appears 11 times.
- 1 `"Captain's name"` appears 10 times.

Distribution du nombre d'apparitions des différents `[Captain's name]` dans `Slave Voyages`.



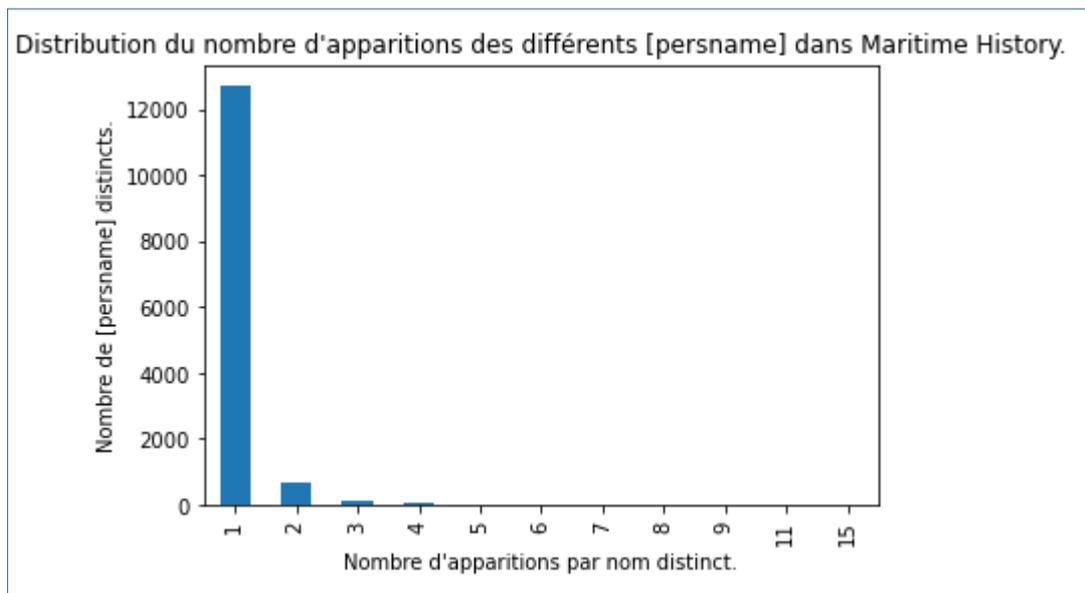
3 – Méthodologie

Les capitaines : exact match.

Il y a 14837 lignes au tableau `Maritime History`, avec 13593 `"persname"` différents.

Comment lire ce diagramme. De gauche à droite :

- 12724 `"persname"` appear 1 time.
- 654 `"persname"` appear 2 times.
- 136 `"persname"` appear 3 times.
- 39 `"persname"` appear 4 times.
- 24 `"persname"` appear 5 times.
- 7 `"persname"` appear 6 times.
- 4 `"persname"` appear 7 times.
- 2 `"persname"` appear 8 times.
- 1 `"persname"` appears 9 times.
- 1 `"persname"` appears 11 times.
- 1 `"persname"` appears 15 times.

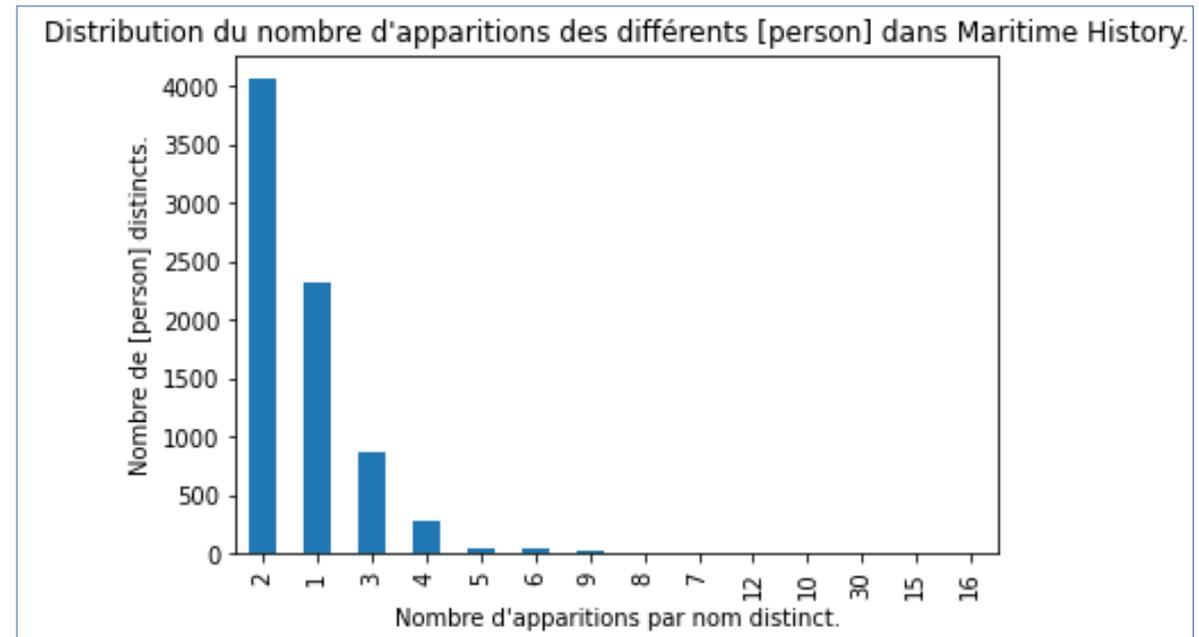


3 – Méthodologie

Les capitaines : exact match.

Il y a 14837 lignes au tableau `Maritime History`, avec 7614 `"person"` différents.
Comment lire ce diagramme. De gauche à droite :

- 4063 `"person"` appear 2 times.
- 2318 `"person"` appear 1 time.
- 857 `"person"` appear 3 times.
- 267 `"person"` appear 4 times.
- 38 `"person"` appear 5 times.
- 36 `"person"` appear 6 times.
- 11 `"person"` appear 9 times.
- 9 `"person"` appear 8 times.
- 4 `"person"` appear 7 times.
- 4 `"person"` appear 12 times.
- 4 `"person"` appear 10 times.
- 1 `"person"` appears 30 times.
- 1 `"person"` appears 15 times.
- 1 `"person"` appears 16 times.



3 – Méthodologie

Les capitaines : exact match.

→ Méthode :

- ↳ Scinder (*split*) les noms sur les espaces. Prendre le tout premier et le tout dernier.
- ↳ Comparaison des « prénoms » entre eux, et des derniers noms entre eux. Si les deux correspondent exactement, on garde.
- ↳ Résultat :

```
{'entities': 9525, 'with possible matches': 1557}  
16.346456692913385%
```

3 – Méthodologie

Les capitaines : exact match.

→ **Christoffel Kok** (Slave Voyages)

- ↳ Nom présent 4 fois dans les voyages 10 391 à 10 394.
- ↳ Les voyages :
 - ↳ 1760-05-15 à 1762-02-15
 - ↳ 1762-05-15 à 1763-08-01
 - ↳ 1763-12-19 à 1765-06-22
 - ↳ 1765-11-15 à 1767-02-15
- ↳ Toujours le même navire : Adriana Petronella
- ↳ Toujours les mêmes armateurs : Jan Swart et fils.

→ Probablement la même personne.

→ Maritime History :

- ↳ <http://geovistory.org/resource/i88689>
 - ↳ Christoffel Jan de Cock/Kok
 - ↳ Départ le 1745-08-12
 - ↳ Navire : Borssele
 - ↳ Armateur : VOC Chamber Zeeland
- ↳ <http://geovistory.org/resource/i89697>
 - ↳ Christoffel Jan de Cock/Kok
 - ↳ Départ le 1761-01-09
 - ↳ Navire : Hof D'Uno
 - ↳ Armateur : VOC Chamber Zeeland

3 – Méthodologie

Les capitaines : exact match.

→ **Jan Pietersz** (Maritime History)

↳ Nom présent 1 fois :

<http://geovistory.org/resource/i89995>.

↳ Le voyage :

↳ 1710-01 à 1711-12

↳ Navire : *Shonauwen*

↳ Armateur : *VOC Chamber Zeeland*

→ Slave Voyages :

↳ **Voyage id. 11 803**

↳ Jan Pietersz

↳ Départ le 1659-08-18

↳ Navire : *Nieuw Westindisch Huis*

↳ Armateur : *West-Indische Compagnie*

↳ **Voyage id. 10 476**

↳ Jan Pietersz

↳ Départ le 1747-01

↳ Navire : *Catharina Galei*

↳ Armateur : *West-Indische Compagnie*

3 – Méthodologie

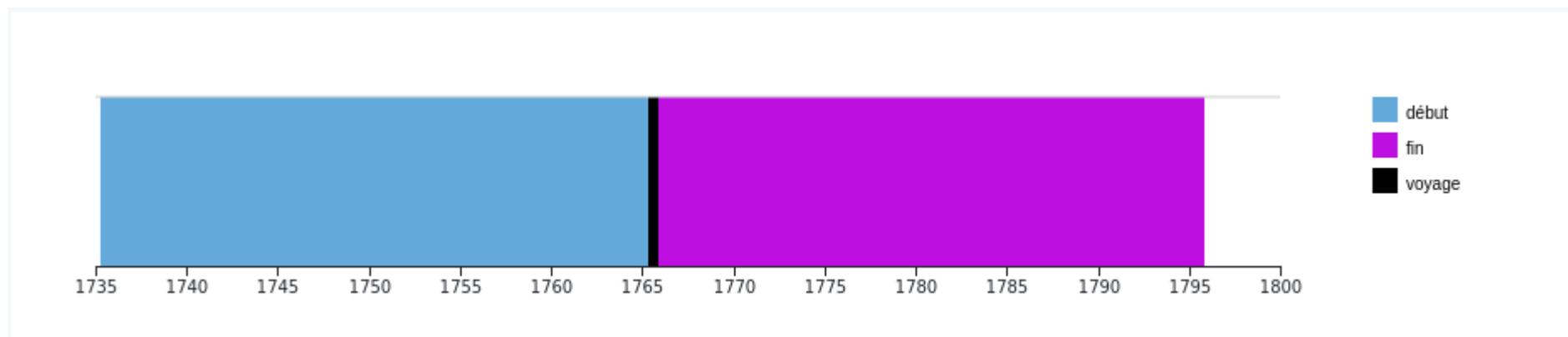
Les capitaines : exact match + dates.

- Pas de date = pas de match possible.
- Estimation arbitraire de la longueur maximum d'une carrière de capitaine à l'époque moderne (30 ans). Pour chaque voyage, la personne pouvant être en début ou fin de carrière, nous avons un gros travail à faire :
 - ↳ Toutes les dates ne sont pas les mêmes. On les répertorie toutes, on les type correctement, on les range dans l'ordre chronologique et on garde la plus ancienne et la plus récente.
 - ↳ On enlève 30 ans à la plus ancienne et on ajoute 30 ans à la plus récente.

3 – Méthodologie

Les capitaines : exact match + dates.

- Maritime History, <http://geovistory.org/resource/i85834>
- Dirk Jansen Backer ; Dirk Jansse Bakker ; Dirk Jansz Backer.
- Voyage entre 1765-04-20 et 1765-11-16.



3 – Méthodologie

Les capitaines : exact match + dates.

→ Liste des conditions :

- ↳ Les prénoms sont identiques.
- ↳ Les derniers noms sont identiques.
- ↳ Le voyage de l'un rentre dans les dates « possibles » de l'autre.
- ↳ Les deux voyages ne se superposent pas.

```
{'entities': 9525, 'with possible matches': 1557}  
16.346456692913385%
```

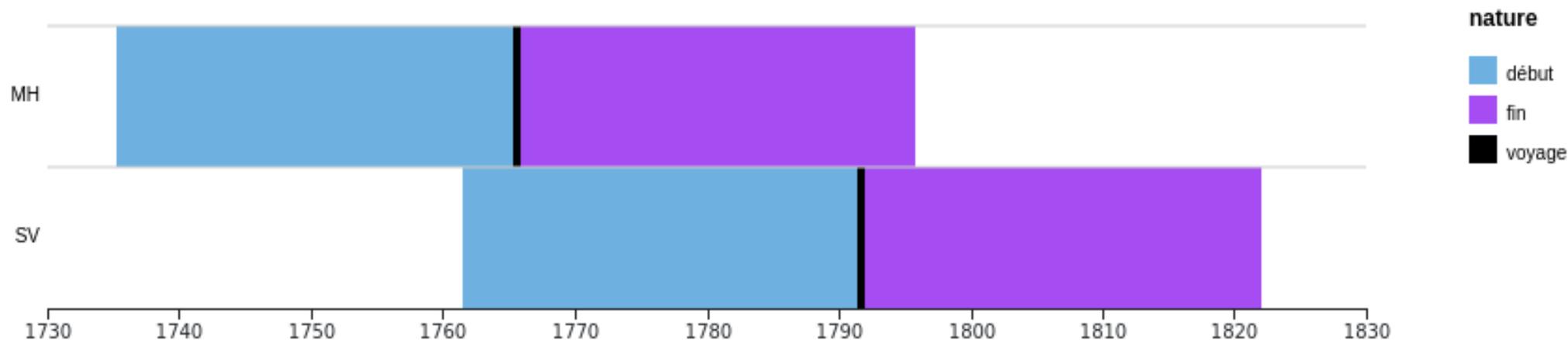
```
{'entities': 9525, 'with possible matches': 1142}  
11.989501312335959%
```

3 – Méthodologie

Les capitaines : exact match + dates.

→ Maritime History, <http://geovistory.org/resource/i85834>, Dirk Jansen Backer ; Dirk Jansse Bakker ; Dirk Jansz Backer.

→ Slave Voyages, voyage id. n° 10 749, Dirk Bakker.



4 – Méthodologie

En conclusion :

- On ne peut pas être sûrs que deux personnes sont les mêmes. Par contre, on peut écarter celles pour lesquelles c'est impossible. C'est à cela que sert le travail présenté aujourd'hui.
- Chaque agrégation est unique. Si l'opération nous épargne le dépouillement des sources, lorsque les données ne sont pas interopérables, l'opération est un casse-tête.
- Pour prévenir les problèmes futurs, il faut choisir des standards internationaux et respecter les principes FAIR.

5 – Discussion/questions

Merci pour votre attention !